The multimodal sentiment analysis in car reviews (MuSe-CaR) dataset: collection, insights and improvements

Lukas Stappen, Alice Baird, Lea Schumann, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Stappen, Lukas, Alice Baird, Lea Schumann, and Björn Schuller. 2023. "The multimodal sentiment analysis in car reviews (MuSe-CaR) dataset: collection, insights and improvements." *IEEE Transactions on Affective Computing* 14 (2): 1334–50. https://doi.org/10.1109/TAFFC.2021.3097002.

The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements

Lukas Stappen, *Member, IEEE*, Alice Baird, *Member, IEEE*, Lea Schumann, and Björn Schuller, *Fellow, IEEE*

Abstract—Truly real-life data presents a strong, but exciting challenge for sentiment and emotion research. The high variety of possible 'in-the-wild' properties makes large datasets such as these indispensable with respect to building robust machine learning models. A sufficient quantity of data covering a deep variety in the challenges of each modality to force the exploratory analysis of the interplay of all modalities has not yet been made available in this context. In this contribution, we present MuSe-CaR, a first of its kind multimodal dataset. The data is publicly available as it recently served as the testing bed for the 1st Multimodal Sentiment Analysis Challenge, and focused on the tasks of emotion, emotion-target engagement, and trustworthiness recognition by means of comprehensively integrating the audio-visual and language modalities. Furthermore, we give a thorough overview of the dataset in terms of collection and annotation, including annotation tiers not used in this year's MuSe 2020. In addition, for one of the sub-challenges – predicting the level of trustworthiness – no participant outperformed the baseline model, and so we propose a simple, but highly efficient Multi-Head-Attention network that exceeds using multimodal fusion the baseline by around 0.2 CCC (almost 50 percent improvement).

Index Terms—Sentiment analysis, affective computing, database, mutlimedia retrieval, trustworthiness

1 Introduction

Global video traffic is estimated to grow four-fold in the coming years [1], accounting for 80 percent of all online traffic in 2019 [2]. On social media, users view eight billion videos daily on Facebook [3] and YouTube has become the second biggest social network with nearly two billion active users and one billion hours watched each day [4]. The internet has undergone a rapid transformation from a largely text-based Web 2.0 to a multimedia, user content-driven net. However, extracting, processing, and analysing relevant information from the huge amounts of semi-structured user-generated data available remains a challenge [5].

Text-based sentiment analysis is now widely used, e.g., for brand perception or customer satisfaction assessment, as machine learning approaches are able to learn rich text representations from data that can be applied to sentiment classification [6], [7]. However, the increased availability of other modalities (e.g., facial and vocal cues) offers new opportunities for affective computing by incorporating diverse information. Fused representations from text and images have shown improvements over unimodal models for the prediction of sentiment and emotion [8]. As well as this, inter-modality dynamics are harnessed through the integration of multiple modalities, and this has brought forward advancements with respect to sentiment prediction [9], [10], [11].

Furthermore, the engagement of multimodal data for sentiment analysis has received an increasing level of attention lately [12]. The interest of the research community and industry to develop methods, for areas including multimodal sentiment analysis – which analyse the interaction between users' emotions and topics in multimedia content – has grown with the wide dissemination of multimodal user-generated content [9], [13], [14]. While it is established that multimodal approaches lead to higher quality prediction results as compared to unimodal input data [15], due to a lack of robustness, it is still an on-going challenge to develop and employ these techniques in real-world applications. Despite the recent progress for constructing larger datasets [16] to explore and develop counter strategies for novel 'in-the-wild' paradigms, there are many areas which remain unexplored to this day.

In this work, we present in detail the process for the collection and annotation of the multimodal sentiment analysis in Car Reviews (MuSe-CaR) dataset.* The MuSe-CaR dataset is a large, extensively annotated multimodal (video, audio, and text) dataset that has been gathered under real world

- *Lukas Stappen, Alice Baird, and Lea Schumann are with EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany. E-mail: stappen@ieee.org, alice.baird@informatik.uni-augsburg.de, lea.schumann@student.uni-augsburg.de.*
- *Björn Schuller is with EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with GLAM – Group on Language, Audio, and Music, Imperial College, London SW7 2BX, London, U.K. E-mail: schuller@ieee.org.*

(Corresponding author: Lukas Stappen.)

Digital Object Identifier no. 10.1109/TAFFC.2021.3097002

* The data are available online: https://www.muse-challenge.org

conditions with the intention of developing appropriate methods and further understanding of multimodal sentiment analysis 'in-the-wild'. To the best of the authors' knowledge, it is more than three times larger than any other continuously annotated dataset aimed at pushing the understanding of multimodal sentiment beyond discrete modelling. Further to this, MuSe-CaR provides never before seen annotations, which explicitly allows for modelling of speaker topic and physical entity in relation to continuous emotions. A selection of the MuSe-CaR dataset was utilised for the First International Multimodal Sentiment Analysis in Real-life Media Challenge (MuSe 2020) [17], which was held at the ACM Multimedia 2020 conference.

In previous research, multimodal sentiment analysis and emotion recognition are often applied in the context of product reviews [18], [19], and sourced from the open-accessible video platform YouTube [18]. With MuSe-CaR, we are influenced by this collection strategy, and have designed the dataset with an abundance of computational tasks in mind. Furthermore, the dominant focus for MuSe-CaR is to aid in machine understanding of how positive and negative sentiment as well as emotional arousal is linked to an entity and aspects in a review (and other user-generated content in general). In doing so, MuSe-CaR aims to bridge fields within affective computing, which currently utilise a variety of emotionally annotated signals (dimensional and categorical).

We collected over 40 hours of user-generated video material with more than 350 reviews and 70 host speakers (as well as 20 overdubbed narrators) from YouTube. The extensive annotations consist of 15 different annotation tiers/tracks (3 continuous dimensional, 3 partially continuous binary, 5 categorical, and 4 automatically extracted tiers). Among others, MuSe-CaR offers conversational topic labelling, the novel continuous dimension of *Trustworthiness*, and full word-aligned transcription.

When selecting the data, it was of particular importance to balance between the uncontrollable 'in-the-wild' influences and constraining properties to allow for meaningful learning and generalisation using current deep learning methods. Such 'in-the-wild' characteristics of MuSe-CaR include: *i) video*: face-angle, shot size, camera motion, reviewer visibility, reviewer face occlusion' ((sun-)glasses), and highly varied backgrounds within a single video; *ii) audio*: ambient noises, narrator and host diarisation, diverse microphone types, and speaker locations; *iii) text*: colloquialisms, and domain-specific terms. However, the contextual interaction with emotions, e.g., towards entities and aspects, is content dependent. The number of different entities, topics, and aspects that appear in the videos requires special consideration in order to create balanced records for supervised tasks based on vision (image object detection), audio, and linguistics (aspect detection) and their derived fusion. For this reason, we have limited the dataset to vehicle reviews. We consider that reviews from the Automobile domain bring together many of these aspects that are lacking in other review domains (e.g., travel, clothing, electronics). For example, the automobile review videos show half-covered faces, different soundscapes (e.g., engine, wind) simultaneously as the reviewers articulate product descriptions and their opinions. Furthermore, most of the reviewers are semi- or professional reviewers (e.g., YouTube channels, influencers). This has several practical advantages: on the one hand, it increases the video quality significantly, on the other hand, it makes the videos more consistent (a broad but similar range of topics around the vehicle is covered, e.g., vehicle safety). Finally, approval to use multiple videos is greatly simplified, as a channel with multiple videos (and independent speakers) only needs to be contacted once and not per video.*

Based upon this dataset, our contributions in this work are:

- First, we give in-depth information regarding the MuSe-CaR selection, collection, and annotation process, which was not addressed in the MuSe 2020 challenge baseline paper [17]. This includes the presentation of additional annotation tiers that have not been utilised and introduced, yet (cf. Section 3.5). We expect that this will assist future participants and other researchers to conduct and interpret studies on MuSe-CaR more easily.
- Second, we revisit the tasks introduced in MuSe 2020 and demonstrate that the limits in terms of performance in, at least one of these tasks, has not been reached, yet. When doing so, by proposing a simple, yet efficient model which utilises state-of-the-art components we beat the baseline of MuSe-Trust by around 50 percent. Furthermore, we describe extensive experiments run to identify the key settings in modelling the novel task of Trustworthiness.

Given the novelty of the described dataset, and the extensive analysis that has already been made by the authors, there are numerous ideas and further directions which could be taken by researchers in the community. With this in mind, we outline an array of research directions which make use of various combinations of the data annotation tiers.

2 Related Resources

In the following, we highlight important databases for MuSe-CaR to build upon which focus on computational sentiment and emotion analysis from audio-visual recordings. For an overview of databases which utilise only one or other combinations of modalities in this area, the reader is referred to recent survey studies (e.g., for text [20], for vision [21], and for audio [22]). An overview of findings based on our specific areas of interest are summarised in Table 1.

2.1 Multimodal Sentiment Analysis Datasets

It is generally accepted [39], [40] that (multimodal) sentiment analysis consists of a holder and the object (subject, entity) that the emotion is evoked from. Furthermore, the survey of [12] divides the field into three major groups: multimodal sentiment analysis i) in (monologue) video reviews e.g., from video platforms [18], [32], [41], ii) in human-machine and human-human interactions [16], iii) analysing of general multimedia content (e.g., images, gifs) from social media [42], and stresses the need for additional datasets to

* We contacted the creators for consent, see Section 3 for details.

TABLE 1
Selection of Multimodal Sentiment Analysis and Affective Computing Datasets Focusing on at Least One of Three Types of Prediction Targets: (Sentiment) Classes, (Primitive) Emotions, Object-of-Interest

Name	Modal	Language	AnDu	# Anno	Sentiment		Primitive			OoI
					# Sent	# Emo	Class	# IncSt	Cont	
MuSe-CAR	V,A,L	EN	40:12	5	5*	×	V,A,T	×	✓	✓
Affective Computing										
Aff-Wild [23]	V,A	EN	‡30:00	6	×	6	V,A	×	✓	×
SEWA [16]	V,A	MULTI 1	4:39	5	×	×	V,A,L	×	✓	×
HUMAINE [24]	V,A	EN	4:11	6	×	×	V,A,I	7	✓	×
RECOLA [25]	V,A	FR	3:50	6	×	×	V,A	9	✓	×
AFEW-VA [26]	V,A	EN	2:28	×	×	×	V,A	21	×	×
VAM [27]	V,A	EN	12:00	6-8	×	5	V,A	5	×	×
IEMOCAP [28]	V,A,L	EN	11:28	5	×	9	V,A,D	5	×	×
SEMAINE [29]	V,A	EN	6:30	6	×	7	V,A,I,P,E	×	✓	×
Belfast [30]	V,A	EN	3:57	6	×	×	V,A	3	×	×
Multimodal Sentiment										
UR-FUNNY [31]	V,A,L	EN	90:23	2	×	1	×	×	×	×
MOSEAS [32]	V,A,L	MULTI 2	68:49	3	7	6	×	va	×	×
MOSEI [33]	V,A,L	EN	65:53	3	7	6	×	×	×	×
ICT-MMMO [18]	V,A,L	EN	13:58	2	5	×	×	×	×	×
Ext. POM [34]	V,A,L	EN	15:40	1	5	×	×	×	×	×
CH-SIMS [35]	V,A,L	CN	2:20	5	5	×	×	×	×	×
AMMER [36]	V,A,L	DE	1:18	1	×	5	V,A	11	×	×
Youtubean [37]	V,A,L	EN	1:11	2	3	×	×	×	×	×
MOUD [38]	V,A,L	ES	0:59	2	3	×	×	×	×	×
YouTube [19]	V,A,L	EN	0:29	3	3	×	×	×	×	×

__Modal__ities available; Language: MULTI 1 = CN, DE, EN, GR, HU, SE & MULTI 2 = DE, ES, FR, PT; __An__notation __Du__ration (hh:mm)†; __#__ Number of minimum __Anno__tations per target. Subjectivity includes Sentiment: __#__ Number of __sent__iment classes (derived sentiment intensity from Valence); __#__ number of (basic) __Emo__tions; __Cont__ionous, Primitive: __Dim__ensions: __V__alence, __A__rousal, __T__rustworthiness, __L__ikability, __I__ntensity, __P__ower, __E__xectiation, __D__ominance; __#__ number of __Inc__rement __St__eps, __T__race annotations; __O__bject-__o__f-__I__nterest: classes of topics or entities. ‡Note: In the case of multimedia sentiment analysis databases, we have usually indicated the total size. However, the exclusion of non-spoken parts typically reduces the size significantly, and authors tend not to specify the size of the adjusted audio-visual database. For example, 33 hours of audio-visual data were collected for the SEWA database, but only 14 percent of the corpus are annotated with emotions. For Aff-Wild, only the total size of the dataset but the extend of annotation is not explicitly stated.*

extend the field. Recently, *UR-FUNNY* [31] collected a large number of Ted-Talk videos on more than 400 topics. On a small part of this data, binary humour is predicted. *MOSEAS* [32] is a multilingual collection of 40 000 audio-visual sentences consisting of sentiment, emotion, and attribute labels. *MOSEI* [33] contains videos of 250 topics for sentiment analysis (7 classes) and emotion recognition (6 classes). Videos with no transcriptions and punctuation are provided by the creator, additionally they exclude video content where the camera was not fixed in place. [43] extended *POM* [44], an audio-visual film review dataset, with annotations at the level of opinion-forming segments and components. It consists of 600 videos with an average length of 94 seconds in which a person looks straight into the camera and talks about six film aspects. *ICT-MMMO* [18] consists of user-generated review videos to predict the sentiment. *CH-SIMS* [35] acquired 60 Mandarin raw videos from movies, and television series and shows with segments of up to 10 seconds. It is limited to parts where both, the face and the voice appears at the same time. *Youtubean* [37] collects seven popular product reviews of one cell phone model with aspect and sentiment annotations. *MOUD* [38] includes *YouTubers* clearly visible (no occlusions) and oriented frontally to the camera expressing their opinions in 30 second segments without any background music. The *YouTube* corpus [19] provides sentiment labels for a wide range of YouTube product reviews. *AMMER* [36] focuses on emotional interactions in a simulated car journey in German. Discrete emotions used in these databases are not fully eligible to represent the sentiment, making more complex representations [12] e.g., through polarity and intensity scores necessary.

2.2 Affective Computing Datasets

One such way to represent the sentiment more comprehensively is through the primitive dimensions of emotions, e.g., the circumplex model of emotions [45]. In this context, Valence often serves as an umbrella term of sentiment and is used interchangeably [46], [47], [48]. *Aff-Wild* [23] consists of 30 hours of audio-video material on emotion recognition (valence, arousal) focusing on facial images in a variety of head poses, lighting conditions and occlusions. It was later expanded [49] by an additional 260 YouTube videos with a total length of about 13 hours, annotated under similiar conditions and at frame level adding action units and 7 basic expressions. *SEWA* [16] provides a large audio-visual data with continuous arousal, valence, and likeability traces during human-human interactions recorded online via static webcams. However, only around 4 hours are continuously annotated. *HUMAINE* [24] provides continuous intensity annotations in addition to Arousal and Valence. The *RECOLA* dataset [25] contains subjects interacting in a tightly controlled laboratory environment. The audio, visual, and electro-dermal activity were annotated with Arousal and Valence traces. *AFEW-VA* includes 7 facial expressions from films, annotated by 3 raters, in addition to Arousal and Valence with intensities from -10 to 10 hard incremental steps. *VAM* consists of clips from a German talk-show. Besides Valence, Arousal, and dominance, which are annotated on a discretised 5-point scale, there are also six basic emotions which are sparsely labelled. *IEMOCAP* [28] consists of audio, video, and transcriptions of ten actors in dyadic sessions in a controlled setting. *SEMAINE* [29] is a richly annotated database of 21 human-agent interaction sessions. *Belfast* [30] is a collection of actively stimulated participants evoke moderate emotional response due to specific tasks. The recordings come with continuous values of Valence and Arousal annotations.

2.3 Summary

From this overview of recent literature, on the one hand, we find that multimodal sentiment analysis databases currently try to select content from a wide range of topics, which increases the (linguistic) generalisation potential of developed models. However, multimedia databases which utilise topics or aspects as a prediction target [37], [43] are rare, limiting the (supervised) understanding of the context almost exclusively to linguistic analysis. To date, no such methods offer complete solutions and rely on the spoken word. In situations where spoken language is not present and towards multimodal understanding, we want to examine the relationships between emotion and object/ physical entity. As the literature suggests a) these topic definitions, e.g., reviews [33] are too high-level to justify an in-depth *understanding* of the opinion-topic structure and their multimodal interactions, a point which is necessary for real-life applications, and b) with the drastic improvements of general language models ([50]) generalisation improves naturally over time [51].

On the other hand, with a wide range of (health and well-being) situations in mind, affective computing focuses on the elementary sensing of affects using often primitive and generalistic (continuous) emotional dimensions, making it very broadly applicable [52]. These enable short- and long-range understanding and aggregation of affects, emotions, and sentiments. However, understanding what these are aimed at (e.g., subject, entity) is not necessarily seen as part of the field. Moreover, the shift towards perception in a noisy environment (in-the-wild), especially with regard to visual characteristics, has only recently been made a focus [16], [26]. We argue that a) with the improvements of models through deep learning and its efficient utilisation on large-scale data, this focus might change, and b) the properties of continues traces might be helpful when dynamically breaking down a large sequence of audio-visual emotional events into shorter segments, e.g., sentences, aspects or noun-adjective pairs in the future.

The MuSe-CaR dataset described in this work was designed to overcome some of these basic limitations when utilising user-generated, real-life media on a large-scale and attempts to bring the best of two worlds together. We consider the following aspects in the design of MuSe-CaR listed below.

- The recording of the database should be as uncontrolled to a high degree in terms of recording settings, emotional and linguistic content as possible. However, there should be a certain overlap in the topics and aspects addressed.
- Instead of isolated segments and sentences, the database should contain long sequences across multiple aspects of a topic.
- The emotion-object interaction of the speaker should be in various environments (e.g., outside and inside of a car).
- The depicted speaker is not actively acting emotions, but rather, emotions are naturally elicited depending on the topic, aspect and situation.
- The audio-visual material provides verbal and non-verbal information. However, occasionally, one modality provides only limited information (voice but no face, e.g., the camera focuses on an object; face but no voice, e.g., acceleration; audio-visual but no transcription, e.g., speech-to-text failed due to complex audio scenarios).
- The emotional and topic annotations should be assigned based on human subjective evaluations.

3 The MuSe-CaR Database

In the following, we describe how we identified relevant material based on the previously defined criteria. Furthermore, we outline the communication process with the creators to reach consent for the use of their content. Then, we provide information regarding the dataset composition, as well as the data annotation process and the utilised annotation tools.

3.1 Data Acquisition

The selection of videos from YouTube was carried out in a semi-automatic process. We developed a basic crawler which receives a number of hand selected keywords (e.g., 'review' and a car brand) and provides metadata of pre-selected videos. In our view, the legal situation for crawling videos from the web is inconclusive in many countries.[§] Therefore, we contacted the creators of the videos with high user engagement (views, likes, etc.) which most likely indicates relevance and high-quality content. Requesting actively the creators' consent to use the data for academic purposes in an opt-in approach, gives researchers world-wide legal certainty when using the database. We sent up to three (follow-up) emails to the creators over a period of three months, reaching an agreement in around 50 percent of requests. In the emails, we explained the intention of the dataset for non-commercial use in challenges and for

§ Uploading a video to YouTube automatically issues that video under the YouTube own license. Regarding this licence, the use of the data in the EU is only possible by YouTube directly or with the consent of the creator. In similar works [13], [32], [33], the database producers refer to the fair use principle for academic use. These exceptions of intellectual property rights, however, do not seem to be applicable in the European legal sphere. Furthermore, YouTube's standard terms and conditions as well as of those from the API have to be considered. A fraction of videos are also available under the Creative Commons (CC-BY, full use if the creator credits are mentioned) licence model.

Fig. 1. Thumbnails showing reviewers in various constellations to the camera and interacting with the object.

research and provided an example of an End User License Agreement (EULA).

3.2 Data Inspection and Selection

If the consent was given, three individuals carried out a deeper initial inspection by viewing around 10 percent of each video (one assessment per video). In total, 366 videos were inspected. Based on the criteria defined in Section 2.3 and the content of each video, the inspectors filled out a survey, asking to estimate important data properties such as the level of in-the-wild characteristics. These also included the emotionality and quality of the video for our purpose (cf. meet the derived criteria in Section 2.3) on a scale of 1 to 5 (where 1 is substandard, and 5 is optimal). For example, a substandard video would rather have shots of a car but is not a review or the majority of the filming took place in a studio. Regarding emotionality, less than 5 percent are rated below 3 and around 80 percent as 4 or 5. Furthermore, regarding video quality properties, they seem suitable for our purpose with more than 85 percent holding scores equal to or higher than 4. For instance, purely neutral videos with no emotional interaction, e.g., without a human, synthetic speech narration listing criteria of the car. Furthermore, additional estimation questions reveal more detailed information about the videos (e.g., face inclusion): shot range, or camera selfie-angles, the scene and noise settings (including the quality of sound), whether there are background music and other speakers present, and if speakers have a dialect or accent. These additional estimation questions, are specifically general, and the annotators prior to observing did not receive detailed instruction on their definition, other than being given the 'emplate' video which was given as a 'gold standard' In regards to nativeness and dialect, although we do not have a ground truth for this, and we notes that particular non-native speakers may have a less obvious accent, we consider consider that deeper analysis that a 'first-impression' may be out of scope.

A detailed analysis can be seen in Fig. 2. Interesting observations are that there are fewer close up than medium sized shots in most videos (only accounts for 1-25 percent of the videos, cf. Fig. 1). For selfie shots, the camera is mainly held at a lower angle. Videos are often filmed inside the car, either while driving or not driving. Fewer coverage accounts for outside-the-car shots (for both driving and not driving). In general, there are noise sounds and background music present, however, the sound quality is only perceived

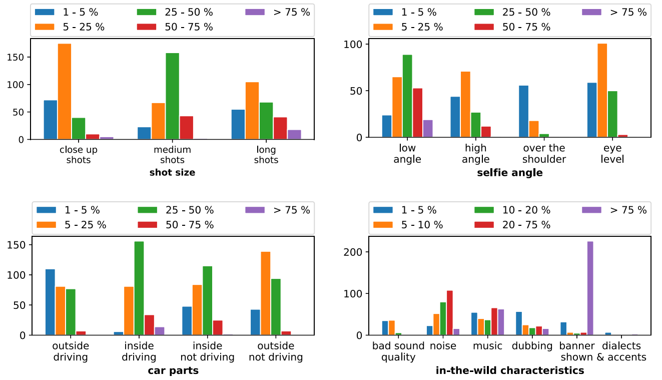

Fig. 2. Absolute counts of the estimation of various in-the-wild characteristics, namely camera shot size, camera selfie-angle, scene setting, and additional noise influences of the collection. The percentage estimates the duration of collected video in a certain category, e. g. , 25-50 percent in "close up shots" means that the collector estimates that 25-50 percent of the duration of the collected video contains close up shots.

to be bad for 1-10 percent of the video duration. The great majority of videos has a banner present (e.g., copyright sign in a corner) for over 75 percent of the video duration.

Since the subjects were not actively selected, but professional, semi-professional ('influencers'), and casual reviewers, we can only estimate the characteristics of our cohort. We assume a broad age range from the mid-20s until the late-50s while most speakers are English natives from the United Kingdom or the United States of America; a small minority are non-native, yet fluent English speakers. Around a third of reviewers wear glasses. There are barely any videos with speakers having a dialect or accent.

Videos which received less than 4 points of overall quality for training, less than 3 on emotionality or lack of the properties outlined in the previous section were excluded. After inspection, 303 videos remained for annotation, which corresponds to 40.2 hours of video with an average duration of 8 minutes (90 percent are shorter than 14 minutes).

3.3 Post Processing: Voice-Activity, Transcription

It is well established that linguistic information of the spoken language can facilitate the learning of emotions and is a cornerstone for understanding context. To support future research into the interplay of the audio, visual, and text modalities, we automatically transcribed the data. In recent years, speech-to-text achieved almost human-level quality in popular languages, such as English. Furthermore, using the text modality for emotion recognition in a real-life scenario would need to work independently, without human intervention, and considering the size of the dataset along with these prerequisites, we decided to use automatic transcription services on our corpus.

The transcriptions from the videos using Google Cloud speech API[¶] and Amazon Transcribe[‖] were both of sufficient quality. The first contains also non-verbal cues and audio elements, such as laughter, music, theme, etc.

Ten videos were randomly selected for human transcription to estimate the word-error rate (WER). The total number of transcribed words is 10 576. The WER of Google

¶ https://cloud.google.com/speech-to-text
‖ https://aws.amazon.com/transcribe/

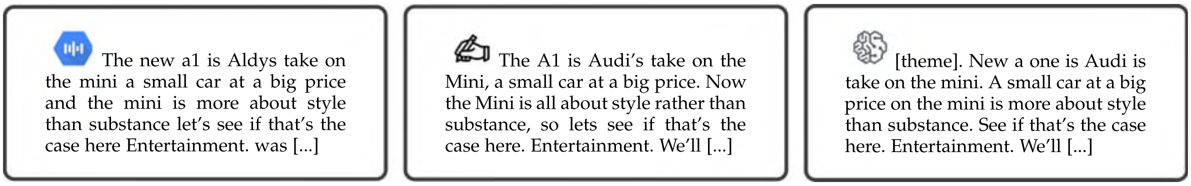

Fig. 3. The video (ID: 265) with the worst speech-to-text result from our hand-transcribed (middle) selection with a word error rate of 37.85 percent from Google-Transcribe (left) and 39.44 percent from AWS (right).

Cloud speech API in this sample corresponds to 25.04 percent, while that of Amazon Transcribe is 28.39 percent. In Fig. 3 we have shown an example of a section from the video with the highest error rate from our sample. This analysis indicates that even if the values seem objectively high, typical errors are often minor such as "A1 is Audi's" (hand) versus "a1 is Aldys" (google) versus "a one is Audi" (AWS). We provide both, but only use the latter for our experiments since we felt that the word timestamps and car-specific vocabulary slightly exceeds the first one. One reason for this improvement may be the option to create a customised dictionary to improve the transcription quality regarding domain-specific, automotive typical terms.

The transcriptions of the spoken language includes punctuation (e.g., period, question mark, exclamation mark) and every transcribed word comes with a beginning and an end timestamp as well as duration. These metadata help to align the text with the annotations (different sampling rate) and other modalities as well as enabling studies on more than 28 295 sentences exceed all English Multimodal Sentiment Analysis databases (cf. Section 2) and have almost 5k more sentences than the next biggest (MOSEI).

Although we have not rigorously evaluated them, our preliminary screening suggests that the punctuation and boundaries are accurate and even better than commonly available voice activity detectors we also have tried.

3.4 Data Annotation

3.4.1 Annotation Roles and Organisation

The size of the dataset and fine-grained annotations required a highly efficient annotation process. In order to ensure high quality, ethical and meaningful annotations for the MuSe-CaR dataset, we considered that keeping a human in the loop was a vital aspect [53]. We therefore defined three functional roles:

1) *Annotator*: It is the responsibility of the annotator to label the data based on the subsequent instruction.
2) *Auditor*: It is the responsibility of the auditor to review the performance of the information labelled, and ensure it is in line with the annotation protocol. Only after the annotations have been manually and automatically checked, verified and endorsed by the auditors is the annotation deemed usable.
3) *Administrator*: Manage all parties, and assign duties during the entire annotation process.

During annotation, an interactive process between the annotator, auditor, and administrator was applied:

1) *Assignment of tasks:* The annotator is assigned one or more packages by the administrator. Similar to previous work [28], a session package corresponds to ca. 40 minutes of video material. The annotators were instructed to have suitable rests between videos and sessions, so that the expected (and paid) working time was one hour. In one session, all videos had to be annotated with the same annotation type (e.g., Valence). The videos are distributed in two rounds. In the first round, three annotators annotate a package and after all of them have finished, round two takes place, where the remaining 2 annotations per package are carried out. This division of allocation is necessary to determine the quality in round one (1 versus 2 annotators) and that no imbalance occurs by improving the quality of late assigned packages. In other words, with equal distribution, early assigned packages have a worse quality than those assigned at the end, see also quality tracking).
2) *Annotation*: The annotator annotates the videos piece by piece and package by package and sends packages to the auditor after completion. Once all packages have been processed and evaluated by the auditor, new packages can be assigned to the annotator.
3) *Progress tracking:* The administrator regularly tracks the progress of the annotators and auditors. This keeps track of which packages are still to be annotated or audited – in the worst case, reevaluating the suitability of the work load.
4) *Quality tracking:* By calculating the similarity (Concordance Correlation Coefficient (CCC)) of the annotations in a "one versus all" strategy of the batch-wise assigned videos, high disagreements can be identified. These are then manually investigated regarding qualitative aspects by the Auditor (see above) according to the annotation protocol. The Auditor provides detailed feedback to the annotators and with continuous auditing throughout, the quality of the annotator can be tracked and the quality is therefore continuously improved.

All annotators hold at least a Bachelor's degree, while the minimum requirement for the auditors and administrators was a Master's degree in a technical field and at least 2 years of work experience related to the research field.

3.4.2 Annotation Tools

For categorical annotation (e.g., speaker topic), we used the annotation software ELAN 4.9.4 [54] – chosen for its multimodal interface which allows for a waveform and video display (see Fig. 4), as well as other useful functionalities including the ability to jump to areas of interest [54].

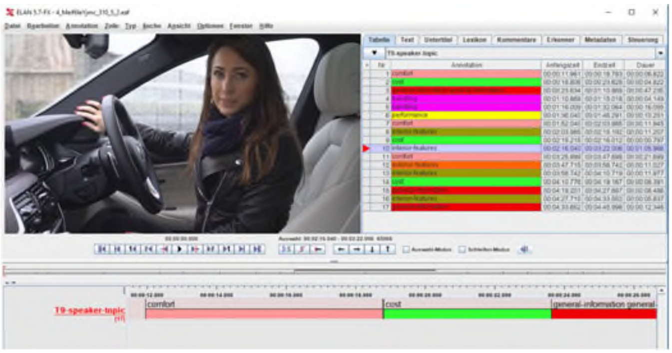

Fig. 4. Interface of the annotation software ELAN when labelling speaker topics (cf. Section 3.5.7).

It was shown that some emotions are transmitted more strongly via visual signals (e.g., sadness) while others more via audio (e.g., anger) [55]. In addition, context information transported by both modalities plays a crucial role in emotion perception [56]. For an audio-visual annotation of the continuous emotions, we choose the software DARMA [57]. DARMA enables to record annotation signals from a *Logitech Extreme 3D Pro Joystick*. The joystick allows for the transfer of perceived emotions more intuitively [16]. The continuous annotation was made from the very start to the end of a video and was sampled at 0.25 Hz with an axis magnitude of range -1 000 and 1 000. In other words, while an annotator was watching a video, the joystick was simultaneously moved in real-time depending on the perceived emotion, i.e., downwards when the valence changes from positive to negative and the joystick signal are recorded by DARMA.

3.5 Annotation Tiers

The MuSe-CaR contains annotations for continuously-valued (Valence, Arousal, and Trustworthiness), binary-valued (host/narrator turns, banner, and person appearance), and categorical (topics, entities) ratings. Overall, we have annotated 11 tiers for each video (cf. Fig. 5). The dimensional annotation reflects the continuous emotional state of the individuals speaking.

In parallel to the continuously-evaluated signals, simple continuous binary (activated/deactivated) annotations are recorded simultaneously by pressing and holding the trigger adjacent to the index finger of the joystick: i) Trustworthiness + the turns between the host and the narrator, ii) Valence + the appearance of banners, and iii) Arousal + the appearance of more than one person.

In addition to the annotations, we provide more than 10 pre-computed features, e.g., facial landmarks, acoustic low-level descriptors (LLDs), hand gestures, head gestures, facial action units, etc., which in other works [11], [16] were also decelerated as (semi-)automatic generated annotations for prediction. A detailed description of these can be found in [17].

Fig. 6. Snapshot of emotional scenes from the MuSe-CaR dataset depicting high Arousal (left) by non-verbally expressing the excitement about the acceleration of the car and verbally positively valenced appreciation of the cars' exterior.

3.5.1 Arousal and Valence

The videos are annotated using a continuous dimensional model of emotion [45], considering both Arousal and Valence (cf. Fig. 6 and Fig. 7). We discuss these together for ease, however, the literature shows that it is important to annotate them separately [16]. In the dataset, an example of high Arousal is *Stressed* or *Elated (happiness)*, however *Stressed* would be negative Valence, and *Elated (happiness)* would be positive Valence.

The annotators were trained in-person. First, the axes were introduced by showing them an explanatory video to impart a general understanding of aspects of emotion. After further explanation and examples, they could experience the handling of the software and the reaction of the joystick in a hands-on session. The test annotations were followed by a group discussion, where individual annotations were compared to an up-hand recorded annotation of an experienced annotator as well as between the group members. The final packages were done individually in a quite environment with headsets.

3.5.2 Trustworthiness

User-generated information has proven useful when creating large (emotional) datasets with real-world content [33].

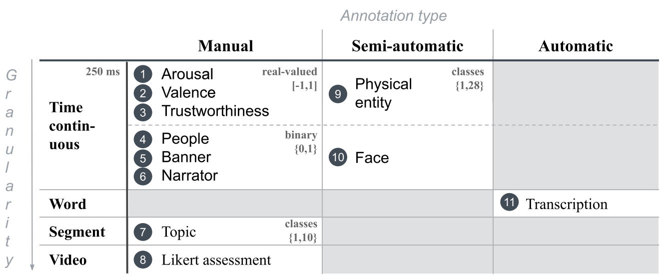

Fig. 5. Overview of the type and granularity (sampling rate) of all our annotations: 1-3: continuous real-valued annotations (cf. Sections 3.5.1 and 3.5.2); 4-5: continuous binary annotations (cf. Sections 3.5.6, 3.5.5, and 3.5.4); 9: physical car entity classes (cf. Section 3.5.8); 10: faces (cf. Section 3.5.9); 11: transcriptions (cf. Section 3.3); 7: speaker topic segments (cf. Section 3.5.7); 8: likert assessment (cf. Section 3.5.10).

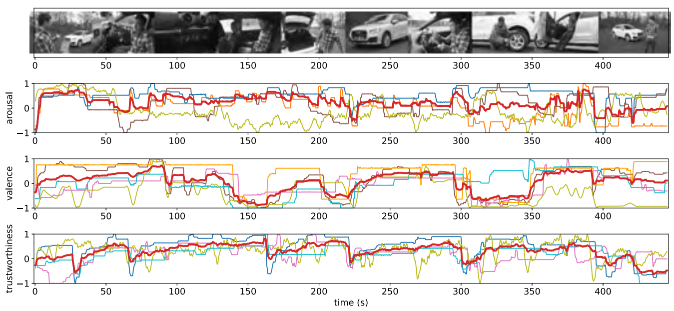

Fig. 7. Annotation example (video id: 236) for the three continuous, real-valued dimensions (valence, arousal, trustworthiness) with a sample rate of 250ms showing the single annotations and the EWE fusion (bold, red).

It is known [58] that to build up reputation of a – in real life unknown – creator has a strong influence to the user engagement and, thus, also most likely on other perceived emotions. However, quantifying Trustworthiness is hard [59], and there is yet no dataset that offers the possibility to link it to Arousal and Valence as well as train cross-domain detectors.

Since this is a completely novel dimension, we explain our definition more deeply. Generally there is no single, prevailing definition of Trustworthiness due to the lack of a conceptual agreement [59], [60], [61]. Analogous to our understanding, in [62], Trustworthiness is defined as the ability, benevolence, and integrity of a trustee.

In the context of a stranger (our moderator) from social media content, Trustworthiness presupposes that this person can objectively assess the (facts of the) matter on the one hand and on the other hand communicates this assessment unbiased, therefore honestly. Building on this definition, we asked the annotators to evaluate the Trustworthiness throughout the video, i.e., when the host is discussing a particular aspect, how honest and knowledgeable does the annotator feel their review is? In other words, this could also include the annotator perceiving a commercial gain rather than a truthful review of the product.

Several video examples and cases were given to the annotators. Although, we reiterate that trustworthiness can be a subjective tier for annotation, and therefore additionally consider this from the annotator's perspective: Do you believe the information that is given to you? Do you have the feeling their argumentation is based on facts and experience or is it that they are rather trying to sell something.

3.5.3 Dimensional Annotation Fusion

Several methods exist to fuse a set of subjective emotion annotations to establish a consensus from individual annotations [63], [64]. Since there can be no fully objective signal of subjective information such as emotions, this label is referred to as the gold-standard.

Fig. 8 depicts the frequency distribution of the created gold-standards utilising the Evaluator Weighted Estimator (EWE) approach [64]. Essentially, EWE considers the reliability of each annotation by calculating the cross-correlation of the annotation with the mean annotation (over all annotators). It can also be seen as the weighted mean of the annotators' agreement [65], [66]. EWE is described further in [64], and has been applied to multiple similar continuous emotion databases [16], [25], [67].

For MuSe-CaR, every video is annotated by at least five independent annotators employed by the EIHW Chair from a group of eleven (six female and five male), all fluent in English. The age of the annotators ranges from 21 to 30 years. In this context, the mean *concordance correlation coefficient (CCC)* is used to measure the inter-rater agreement across all annotations for each dimension is given as: Arousal .265, Valence .350, and Trustworthiness .316. The levels of agreement are moderate as to be expected [68] and consistent with those of other emotional datasets [16], [69] showing that the stronger 'in-the-wild' characteristics seem not to have a strong influence on the perception of emotions. We would like to note that the trustworthiness

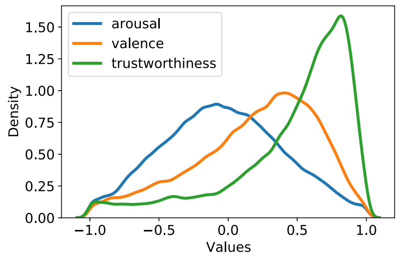

Fig. 8. Density estimation of the continuously-annotated dimensions Arousal, Valence, and Trustworthiness tiers after EWE fusion of individual annotator tracks. While Arousal is almost perfectly Gaussian-shaped, Valence is skewed toward the positive end of the spectrum, and Trustworthiness is even more peaked.

dimension in Fig. 8 is strongly left-skewed. Additional research is required to understand whether the underlying cause of this phenomenon, e.g., our data source (YouTube), domain, or selection approach.

3.5.4 Host & Narrator Turns

Within the videos, there are two human subjects which are of main interest to the annotators; these have been defined as follows:

- The *host* is the reviewer/presenter of the car and its features, talking often directly to the audience and expressing own opinions.
- The *narrator* presents in some videos, may occur in the opening and closing sequences, may provide additional information, and is not visible to the audience.

During the annotation of Trustworthiness, the turns between the *host* and the *narrator* were annotated using the binary trigger. When the host is not speaking, and the narrator speaks, the binary trigger is held.

3.5.5 Banners

Social media networks, such as YouTube, are a large, self-extending and diverse data pool. In these videos, however, superimposed graphics occur such as text and channel logos – some for copyright reasons, others to either inform or entertain the viewer. For our tasks, such banners might not be useful, since they might obscure entities or the host. Therefore, visible banners that show up are annotated in parallel to Valence. In a later stage we want to detect, and measure the influence on visual features or exclude/ replace them [70], [71], [72].

Banners appear with a variety of properties: i) appearance: static, or dynamic, ii) positions: bottom-right, bottom-left, upper-right, upper-left, entire footer, centre, or changing, iii) timing: on-screen billing, opening statement, in-between, or closing credits, iv) duration: highlight (very short), short (few seconds), or consistent (copyright), and v) transparency: none, partly, or mostly transparent.

3.5.6 Multiple People

The focus of this database is on videos with only one visible speaker/ person. In a few cases, more than one person may appear, either for specific comments or for demonstration purposes, e.g., how many people fit in the back seat, or the

TABLE 2
Annotated Speaker Topics With Examples of Sub-Topics and Aspects

Feature exterior (7 %)		Feature interior (6 %)	
light	headlight, foglight, taillight	audio system	radio, speaker
door exterior	locks, handle	seat	belt, split folding backs
Handling, Driving Experience (13%)		**User Experience (7 %)**	
driving actions	braking, steering, gear shifting	infotainment	screen, bluetooth, realtime traffic
dynamics	centroid, chassis, suspension	interaction	interface, iDrive system, gestures
Performance (13 %)		**Quality & Aesthetic (7 %)**	
powertrain	electric, hybrid, combustion	design	interior, exterior, style (sporty, etc.)
engine	horsepower, RPM, acceleration	quality	material quality, clearance
Safety (2 %)		**Comfort (6%)**	
tests	Euro NCAP, NHTSA, rating	surface	leather, touch
assistance sys.	anti-lock brakes, traction control	space	leg room, head room, luggage
General information (16 %)		**Costs (3 %)**	
introduction	series, weight, sales, warranty	one-off	retail price, base price, feature price
comparison	models, brands, competitors	after sale	insurance, maintenance, resale

The % gives the share of the number of sentences in each topic, while 20 percent of sentences have multiple topics.

host is being shown a novel feature. These situations are binary annotated in parallel to Arousal.

3.5.7 Speaker Topic

Speaker Topics rely on generalisable topics vocalised by the host and related to the object of interest. To obtain the topic labels, we started by selecting a broad range of the most mentioned nouns from automatically extracted transcriptions. This list of elements was then reduced manually by watching a random selection of 20 videos from the dataset, gradually gaining knowledge of more frequent topics, based on speaker interests and expressions. These higher-level topics combine many different aspects under one term. The videos are labelled by speaker topic segments while often one topic segment compromises of several sentences (cf. Fig. 4). A detailed overview of sub-topics and aspects covered by each topic can be found in Table 2. It also shows the distribution of the topics across all 28k sentences, while 20 percent have more than one topic annotated.

3.5.8 Physical Entities

In contrast to the voice-based speaker topics, the physical entities rely purely on visual input. The domain-specific objects of human-object interaction, and car parts, were annotated with bounding boxes for all recordings with a step size of 4 frames per second. 28 types of exterior and interior car parts, such as, door, steering wheel, and infotainment were labelled.

Manual bounding box annotation of all frames, each with a high number of classes, are highly labour intensive. Based on previous experiments [73], [74], we can assume that around two boxes per minute can be labelled. Since there are 576 000 frames to annotate with up to 15 boxes for each frame, it would be impractical to label all of them manually (requiring between 4 800 and 72 000 hours for a single annotator).

We chose a semi-automatically process. First, a localiser (Darknet-53 network) is pre-trained on 15 003 vehicle images from other datasets. This underlying data does not include any human interaction. Therefore, we extracted and labelled another 1 000 frames compromising of more than 8 000 boxes from MuSe-CaR and fine-tuned the algorithm. A detailed description can be found in a separate paper [75]. The network achieves a mean average precision of 67.6 percent with scores up to 94.0 percent for very distinctive parts. A visual inspection was performed on the annotations and those deemed unsatisfactory were removed or corrected. Fig. 9 shows examples of these parts during human-object interactions.

Fig. 9. Examples of interaction between the host and the car parts from GO-CarD [75].

3.5.9 Face Extraction

For the same reasons as in Section 3.5.8, we applied a semi-automatic process to the extraction of faces, the start (start of visible face) and end (end of visible face) point of occurrence as well as the relative position within a frame. We labelled faces in a small selection of videos from each channel by hand, to measure quantitatively the success of our automatic extraction and localisation. MTCNN [76] provides a robust framework for this task, previously already proved to be accurate for face annotation in an emotional context [16]. It has a cascaded structure of three stages and is trained on the datasets WIDER FACE [77] and CelebA [78].

We classified the detected bounding boxes into true and false positives resulting in an accuracy of 90 percent, and an F1 score of 86 percent on our selection. Furthermore, we conducted a visual inspection of the bounding boxes. Given the high level of visual in-the-wild characteristics, for instance, partly visible faces, different sizes, side-shots, sunglasses, etc., we consider both, quantitative and qualitative results as strong and sufficient for further feature extraction.

As described in [17] in detail, we use frameworks such as VGGFace [79], and OpenFace [80] to extract face-related features including intensity and presence Facial Action Units (FAUs), facial landamrks, head pose, and gaze position.

3.5.10 Additional Annotation-Related Metadata

Each dimensional and categorical annotation is followed by a very brief survey of the annotator's perception of the content viewed and a self-assessment of their own annotation. On a 10 point Likert scale (0 not at all, 10 very much) annotators are asked four questions: i) How appealing did you find the video? ii) How emotional did you find the host? ii) How trustworthy did you find the content? vi) How confident are you about the accuracy of your annotation?

This data is directly linked to each annotation. An overview of the answers collected is depicted in Fig. 10. Although being somewhat subjective attributes to evaluate, we see a general agreement across our annotations for all questions with a coefficient of variation between 0.12 and 0.25. The appeal of the videos to the annotators seems rather strong with $>$ 20 percent of the ratings between seven and eight. Similarly, the level of emotionality and trustworthiness portrayed by the hosts were perceived high. Regarding the self-assessment of the annotation quality, the annotators seem very confident in their performance.

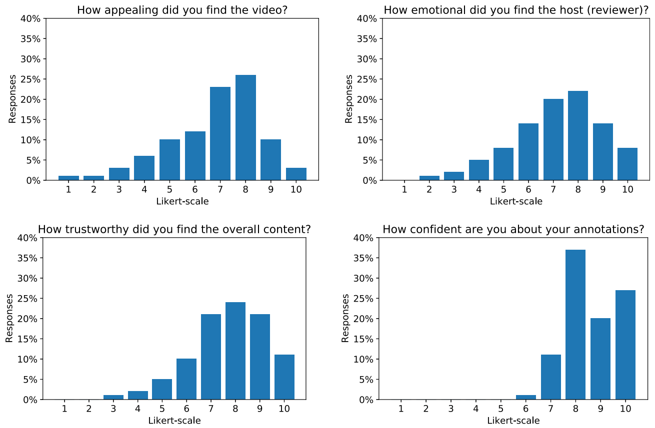

Fig. 10. Annotation related-metadata of almost 5 000 ratings, where the annotator gave scores from 1 to 10 regarding the video watched.

3.6 First Proposed Tasks and Data Availability

For the MuSe 2020 challenge, we proposed a selection of tasks, and a detailed description can be found in [17]. To make the (pre-processed) data for each task (e.g., specific features) easily accessible after the challenge, we moved them to several Zenodo repositories,** a high-speed research data host with storage in the CERN data centre. The MuSe-CaR database is available online for researchers who fulfil the requirements of the EULA (e.g., academic-use only). The specific links can be found in each of the following subsections.

Multimodal Sentiment in-the-Wild (MuSe-Wild).†† MuSe-Wild aims to predict the level of the affective dimensions of Arousal and Valence in a time-continuous manner. Time-stamps to enable modality alignment and fusion on word-, sentence-, and utterance-level as well as several acoustic, visual and textual-based features are pre-computed and provided with the task package.

Multimodal Emotion-Target Engagement (MuSe-Topic).‡‡ The *MuSe-Topic task* focuses on the prediction of 10-classes of domain-specific (automotive, as given by the chosen database) topics as the target of categorical Valence and Arousal emotions. The three classes (low, medium, and high) of Valence and Arousal are each predicted for every topic segment. These classes are created by averaging the mean value of the temporally aggregated continuous labels of MuSe-Wild and then dividing them into three equally sized classes (33 percent) for each label.

Multimodal Trustworthiness (MuSe-Trust).§§ The last task aims to develop methods to predict a continuous Trustworthiness signal in a sequential manner. Aligned Valence and Arousal annotations are also provided, to explore the relationship between all three dimensions e.g., by training multi-task networks.

The evaluation metric of MuSe-Wild and MuSe-Trust is the CCC, which is often used in similar tasks [67], [81] as it stands for a theoretically well understood measure [82], able to reflect the reproducibility and performance while being robust to changes in scale and location [83]. MuSe-Topic is measured in a combination of F1 and Unweighted Average Recall.

** Metadata, and raw material: https://zenodo.org/record/4651164
†† MuSe-Wild : https://zenodo.org/record/4134609
‡‡ MuSe-Topic : https://zenodo.org/record/4134733
§§ MuSe-Trust): https://zenodo.org/record/4134758

TABLE 3
Partitioning of the Three Tasks

Partition	No.	MuSe-Wild	MuSe-Topic	MuSe-Trust
Train	166	22 :16 :43	22 :35 :55	22 :45 :52
Devel.	62	06 :48 :58	06 :49 :46	06 :52 :22
Test	64	06 :02 :20	06 :14 :08	06 :12 :53
$\sum$	291	35 :08 :01	35 :39 :49	35 :51 :07

Reported are the number of unique videos, and the duration for each task in hh: mm:ss.

Table 3 shows the size of the training, validation, and test partitions which consider emotional ratings, speaker/ channel independence, and duration, and come with the packages. In other words, the same host/speaker does not appear across partitions. Before partitioning, we removed less informative data in a pre-processing step. MuSe-Wild and MuSe-Trust only includes parts where an active voice or a visible face are included. Only MuSe-Trust includes non-product related video segments, such as, advertisements which might have an impact on the Trustworthiness perception. MuSe-Topic , the more NLP-related task only includes parts with a speaker topic annotation and transcription.

4 Surpassing the MuSe-Trust Baseline

Although the organisers of MuSe 2020 received several prediction submissions for the MuSe-Trust task, none could surpass the results of the baseline models, resulting in no papers for this task being accepted for the official challenge workshop [84]. In this section, we show that a simple but efficient neural network architecture called DeepTrust utilising a Multi-Attention Head Layer (MAHL) for encoding in addition to a bi-directional (bi) Long Short-Term Memory Recurrent Neural Network (LSTM) with augmentation is suitable for modelling Trustworthiness. We use provided feature sets from the challenge as well as extracting new ones, shown to be effective in the other MuSe tasks, resulting in two different (for acoustic and vision: one handcrafted, and one based on deep representations) feature sets for each modality. We run extensive experiments regarding augmentation, complexity, architecture, and learning the impulse (loss) of such a network. Furthermore, we evaluated the performance of unimodal features and multimodal fusion, as well as the training style (single- versus multi-task learning) for the prediction of Trustworthiness.

4.1 Features

Acoustic. For the handcrafted acoustic feature set, we use the extended Geneva Minimalist Acoustic Parameters Set (eGeMAPS) [85] provided by the MuSe challenge. It is based on 23 acoustic spectral, cepstral, and prosodic low-level descriptors (LLDs) from which statistical functions are calculated. We extract this 88-dimensional feature vector with a window size of 5 seconds and a hop size of 250 ms to enable an alignment to the annotation sampling rate. We further apply standardisation to the vector dimensions.

In addition, we extract VGGISH functions [86] pre-trained on an extensive YouTube audio dataset (AudioSet) [87]. The underlying data contains 600 classes and the recordings contain a variety of 'in-the-wild' noises that we expect to be beneficial to obtain robust features from our 'in-the-wild' videos. By aligning the frame and hope size to the annotation sample rate, we extract a 128-dimensional VGGISH embedding vector every 0.25 s from the underlying log spectrograms.

Vision. The Facial Action Units (FAU), are widely adopted in tasks close to emotion recognition, describing visually perceptible facial movements. FAUs break down facial expressions into 17 individual components of muscle movement, which we obtain from the OpenFace Toolkit [88]. For the deep features, the pre-trained VGGFACE features [79], which were originally developed to identify faces of people are utilised. By removing the last softmax activation layer, we acquire a vector feature representation of the face.

Text. A standard method to transfer words from a symbolic to a continuous representation are word embeddings. These calculate a static, numerical vector per word, depending on the semantics in which the word occurs during training. We extract a 300 dimensional FASTTEXT vector for each word of the automatic transcription [89]. Instead of a static vector representation, context-based NLP transformers extract one vector per word in direct dependence on the context – during the time of inference. For this technique, we apply a BERT Model [90], well established for a number of NLP tasks, and extract the sum of the last four layers as a 768 dimensional feature vector similar to [91].

All features are aligned using the timestamps introduced in Section 3.5.

4.2 DEEPTRUST: Multihead Attention Network for Trustworthiness Prediction

We utilise two neural network mechanisms to model the short- and long-term dependencies of continuous Trustworthiness: enhancing the encoding of the input state by a MHAL and modelling the temporal dynamics of state changes by a LSTM. The attention heads improve the local representation of the extracted features and are able to sustain the long-term (global) dynamics of a sequence in its representation. However, this does not inherit a deeper positional understanding by nature [92]. To this end, we use the functionality of LSTMs, which are particularly capable of learning short- and mid-term patterns. Similar architectures were used by [91] and [93] for emotion recognition.

Mathematically, we train a function $F(X_i)$, where X is a sequence of uni- or multimodal input features, which predicts a sequence of regression point estimates y_i. For this, we apply h multi-attention heads to obtain more meaningful sequence representations s_t, with T being the maximum number of steps. Multi-attention heads are the key building block of transformer networks. In this process, h softmax dot-product attentions (self-attention mechanism, *Att*) are calculated in parallel to learn different discriminative patterns in each head from the three linear projection inputs (query Q, key K, and values V), whereby the division by $\sqrt{d_k}$ prevents very small gradients. After scaling, the results of the individual heads are concatenated and fed into a subsequent linear layer W^S

$$\mathrm{MultiHead} = s_t = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^S, \tag{1}$$

$$\mathrm{head}_h = \alpha\Big(QW_h^Q, KW_h^K, VW_h^V\Big), \tag{2}$$

$$\alpha(\mathrm{Q, K, V}) = \mathrm{softmax}\Big(QK^T/\sqrt{d_k}\Big)V. \tag{3}$$

The resulting enhanced sequence S is the input of, e.g., a one-directional LSTM to receive the temporal encoded sequence O

$$o_t = \overrightarrow{LSTM}(s_t), t \in \{1, \ldots, T\}. \tag{4}$$

Finally, the temporally encoded information are feed into a regression layer, predicting y_i.

4.3 Experimental Settings

Unlike most large emotional datasets with continuous annotations, MuSe-CaR does not break the videos down into artificial, equally-sized segments, so that content and context can remain largely intertwined. However, this has the disadvantage that some sequences end up very long ($>$ 5 000 steps) increasing the amount of computation power needed. To solve this issue and increase the amount of data, the length is segmented to a fixed number of sequence steps ws moving with a hop size hs as proposed by the baseline paper [17] and several participants [91]. Furthermore, we utilise the segment id, as it provides the models with an additional positional encoding [91].

The models are trained for a maximum of 100 epochs using an Adam optimiser, while the learning rate is reduced, as it reaches a plateau for more than 10 epochs. As in the challenge, the CCC is evaluated on the development set after each epoch, and after the training, the best configurations is subsequently evaluated on the test set.

As we aim for an integrated approach of modality fusion, we use early (concatenate the inputs) as well as late (using a LSTM) fusion in order to better learn from the interaction of the modalities.

For the ablation study, we choose to set $ws = 200$, $hs = 100$ after initial experiments, and the number of hidden neurons of the a bidirectional LSTM to 64. If not otherwise stated, we run a hyperparameter optimisation using $h \in \{2, 4, 8\}$ heads, $lr \in \{0, 0001, 0.001, 0.005\}$, a batch size $\in \{512, 1024, 2048\}$ and report the best result indicating the hyperparameter setting by HP.

4.4 Results

Unimodal. First, we compare the performance of our feature sets. As shown in Table 4, the advanced BERT features and the deep acoustic features VGGISH yield higher results in terms of CCC. Only the results on the vision features behave contrary, where the deep face features VGGFACE seem to easily overfit on the development set, while FAU scores are lower, but generalise better. Early fusion of the both text modalities adds a small advantage compared to single use. For all other modalities, the results are worse on the test set. When comparing the model with EGEMAPS and FASTTEXT features with those of the baseline, both show large

TABLE 4
Results of MuSe-Trust Using the Devel(opment) and Test Set, Applying Early and Late Fusion

	Feature sets	Dev.	Test	HP
Text	FastText	.4559	.4782	I
	BERT	.5624	**.5539**	I
	BERT + FastText	**.5648**	.5478	II
Audio	eGeMAPS	.3921	.1220	III
	VGGish	**.5376**	**.4035**	I
	VGGish + eGeMAPS	.4751	.2402	I
Vision	FAU	.3675	**.3623**	I
	VGGFace	**.4000**	.2802	IV
	VGGFace + FAU	.3936	.3298	V

Results are reported in concordance correlation coefficient (CCC). As feature sets for text: FastText (FT) and BERT; acoustic: VGGish and eGeMAPS; vision: VGGFace and FAU for our models.

improvements on the development set and the result for FastText almost doubled on the test set. For the following experiments, we use the best performing feature set of each modality, namely BERT , VGGish , and FAU .

Augmentation. Previous challenges [81] identified the effective use of the available data as a key performance driver. Table 5 shows the results under changing number of sequence steps ws and hop sizes hs. When ws is equal hs, the sequences have no overlap. A larger ws appears to be generally valuable. Most likely, this variable (ws) improves the ability to capture global dynamic changes, leading to a more expressive representation of the state of trust. Having no overlap yields stable, generalisable results, while, when applying an overlap, the results could be either better or worse. As a rule of thumb, the longer the sequences are, the higher the overlap can be, while a good estimate ranges between hs = 0.3 – 0.5 ws. However, if ws is small (e.g., 100, 200), the results might not be generalisable to test. This might be counteracted by applying additional augmentation to the reappearing sequence steps.

Other side effects are that, as the length of the sequence increases, both the memory requirement and the training time grow. With our standard architecture, a maximum of ws = 750 is supported on a 32 GB Memory GPU, which makes it necessary to trade-off performance for usability.

TABLE 5
Results of MuSe-Trust Using the Devel(opment) and Test Set

steps		BERT		VGGish		FAU		Ø	
ws	hs	Dev.	Test	Dev.	Test	Dev.	Test	Dev.	Test
750	750	.5641	.5540	.4274	.4344	.3775	.3719	.4563	.4534
750	500	.5739	**.5747**	.5604	**.4699**	.3671	.3705	.5005	.4717
750	250	.5889	.5693	.5386	.4686	.4305	**.4843**	.5193	**.5074**
200	200	.5512	.5245	.5566	.4752	.3614	.2710	.4897	.4236
200	150	.5500	.5533	.5517	.3034	.3558	.3508	.4858	.4025
200	100	.5624	.5539	.5376	.4035	.3675	.3623	.4892	**.4399**
200	50	.5282	.5160	.5440	.4081	.3319	.1820	.4680	.3687
100	100	.5167	.5128	.5064	.4445	.3711	.3709	.4647	**.4427**
100	50	.5312	.5068	.5369	.2918	.3816	.3294	.4832	.3760
100	25	.5233	.5216	.5264	.2517	.3642	.2911	.4713	.3548

Results are reported in CCC.

TABLE 6
Results of MuSe-Trust Using the Devel(opment) and Test Set

heads	BERT		VGGish		FAU		Ø	
	Dev.	Test	Dev.	Test	Dev.	Test	Dev.	Test
2	.5698	.4745	.5375	.4368	.3561	**.3774**	.4878	.4296
4	.5624	**.5539**	.5376	.4035	.3675	.3591	.4892	**.4388**
8	.5539	.5454	.4035	.2671	.3623	.3280	.4399	.3802
16	.5693	.5112	.5619	**.4592**	.3548	.3352	.4953	.4352

Results are reported in CCC.

Heads. In a unimodal setting, the most suitable number of heads varies from modality to modality, with no clear tendency (e.g., number of dimensions). The best results (cf. Table 6) are obtained with 4 heads for BERT (.5495 on test), 16 heads for VGGish (.4592 on test), and 2 heads for FAU (.3774 on test). On average, 2, 4, and 16 heads perform very similarly on the development set with a slight advantage for 16 heads (.4352) and 8 heads (.4388) on test data.

Loss. Since the loss and the metric are the same (CCC), we also report the Pearson Correlation Coefficient (PCC) and the Root Mean Square Error (RMSE) for this experiment cf. Table 7. The CCC loss clearly performs better than the MSE and L1 loss, for the BERT and VGGish features as well as for the average results of the two correlation based metrics (CCC and PCC). However, this is not the case for FAU where L1 and MSE perform equally or outperform the CCC loss. For RMSE, FAU has such a strong impact that also the average RMSE of both other loss functions are better.

Model. Next, we compare several architectural choices on the unimodal feature selections. As we can see in Table 8, using the combination of both modules is sensible. One configuration (2 MHAL+Bi-LSTM) yields a better result on the BERT test set. For all others and on average, the one-layer MHAL and a bidirectional LSTM architecture achieves the best results.

Multimodal Fusion. Fusing the best single modalities improves all results on the development set. However, the best – the fusion of BERT and VGGish – generalises poorly (devel: .5833 CCC versus test: .5108 CCC) achieving a lower result than BERT only. In comparison, the fusion of text and vision features generalises well and achieves .5880 CCC on

TABLE 7
Results of MuSe-Trust Using the Devel(opment) and Test Set

loss	metric	BERT		VGGish		FAU		Ø	
		Dev.	Test	Dev.	Test	Dev.	Test	Dev.	Test
CCC	CCC	.5624	.5539	.5376	.4035	.3675	.3623	**.4892**	**.4399**
	PCC	.5684	.5998	.5384	.4421	.3770	.4301	**.4946**	**.4907**
	RMSE	.3652	.3485	.3867	.4199	.4693	.4780	.4071	.4155
L1	CCC	.5076	.5211	.3650	.2408	.3678	.3407	.4135	.3675
	PCC	.5432	.5712	.3877	.3270	.3724	.3728	.4344	.4237
	RMSE	.3595	.3409	.4031	.3978	.4350	.3690	.3992	**.3692**
MSE	CCC	.5215	.5433	.3932	.4094	.3537	.3243	.4228	.4257
	PCC	.5455	.5570	.3932	.4100	.3584	.3498	.4324	.4392
	RMSE	.3584	.3470	.3932	.4160	.4407	.3796	**.3974**	.3809

Results are reported in CCC, PCC, and RMSE.

TABLE 8
Results of MuSe-Trust Using the Devel(opment) and Test Set

network	BERT		VGGish		FAU		Ø	
	Dev.	Test	Dev.	Test	Dev.	Test	Dev.	Test
MHAL	.3117	.3248	.4230	.3150	.3677	.3351	.3675	.3250
LSTM	.5165	.5170	.5441	.3771	.3270	.2513	.4625	.3818
MHAL+LSTM	.5423	.5526	.5368	.2248	.3609	.3047	.4800	.3607
MHAL+2 Bi-LSTM	.5456	.5504	.5259	.3688	.3642	.2973	.4786	.4055
MHAL+Bi-LSTM	.5624	.5539	.5376	**.4035**	.3675	**.3623**	.4892	**.4399**
2 MHAL+Bi-LSTM	.5548	**.5762**	.4918	.3818	.3645	.3447	.4704	.4342
2 MHAL+2 Bi-LSTM	.5410	.5344	.4942	.3233	.3553	.3523	.4635	.4033
3 MHAL+3 Bi-LSTM	.5437	.5089	.4977	.3376	.3455	.3104	.4623	.3856

Results are reported in CCC.

test data (.5803 CCC on development) – a better result than every single modality. Similarly, VGGish and FAU achieve slightly higher results on both sets when being fused resulting in .4287 CCC on the test set.

Multi-Task Learning. Using our approach to predict Arousal, Valence, and Trustworthiness simultaneously, outperforms the baseline by more than .2 on development and almost .15 on the test set. (cf. Table 9) Adding more weight to the Trustworthiness predictions (II.) improves the result slightly.

Best Configuration. Our best results are shown Table 10 achieving on development .6507 CCC and on test .6105 CCC (late) fusing all except the VGGish predictions of the fully trained uni-modal models. The results represent a major improvement in terms of the baselines established by [17] more than 50 percent on the development set.

5 Future Work and Limitations

In the future, we plan to extend the first proposed tasks in several directions by using the additional levels of annotation presented. Of particular interest are the possible presence of a connection between the novel dimension of Trustworthiness and Arousal/ Valence and its usability to evaluate user-generated data, which could give insights on the subjective perception of Trustworthiness.

One focus for the MuSe challenge is to bring together communities from differing computational disciplines; mainly, the sentiment analysis community preferring to predict discrete sentiment/emotion categories coming from an NLP background [95], and the audio-visual emotion recognition community predicting continuous-valued Valence and Arousal dimensions of emotion (circumplex model of affect [45]) originated in intelligent audio and visual signal processing, while often disregarding the

TABLE 9
Results of MuSe-Trust Using the Devel(opment) and Test Set

Configuration		T		A	V
Model	Features	Dev.	Test	Dev.	Dev.
End2You-Multitask [17]	FastText + VGGFace + A	3264	.4119	–	–
MHAL+LSTM-Multi (I.)	BERT + VGGish + FAU	.5428	.5456	.4102	.4442
MHAL+LSTM-Multi (II.)	BERT + VGGish + FAU	.5497	.5518	.4132	.4215

*Results of Trustworthiness (T), Arousal (A), and Valence (V) reported in CCC. Configurations: (I.) equal loss weight or 0.33 (II.) 0.5 x Trustworthiness, 0.25 * {Arousal,Valence}.*

TABLE 10
Results of MuSe-Trust Using the Dev(elopment) and Test Set

Model	Features	Dev.	Test
MultiFusion [94]	FastText + DS + 2D	.3426	.3259
Official Baselines [17]			
LSTM+Self-ATT	eGeMAPS	.1576	.1385
LSTM+Self-ATT	FastText	.2278	.2549
LSTM+Self-ATT	FastText + eGeMAPS	.2296	.2054
LSTM+Self-ATT	FastText + eGeMAPS + V	.1245	.1695
End2You	FastText + VGGFace + A	**.3198**	**.4128**
Ours (fusion)			
DeepTrust best-of (early)	BERT + VGGish + FAU	.6241	.5073
DeepTrust (early)	BERT + FastText +VGGish + eGeMAPS + FAU	.5445	.4998
DeepTrust best-of (late)	BERT + VGGish + FAU	.6075	.5796
DeepTrust (late)	BERT + FastText +VGGish + eGeMAPS + FAU	**.6507**	**.6105**

Results are reported in concordance correlation coefficient (CCC). As feature sets for text: FastText (FT) and BERT; acoustic: VGGish, eGeMAPS, and DeepSpectrum (DS); vision: VGGFace, FAU, and 2D landmarks (2D). Furthermore, the baseline models use the raw audios (A) in End2You and several combined vision features (V) in LSTM + Self-Att.

potential of the textual modality [67], [81], [96], [97]. However, both have its advantages, e.g., classes of emotions are more intuitive for humans (happy label versus Arousal and Valence scores) and dimensions are more generaliseable and are highly influenced by related, explicitly multimodal learning, techniques [98], [99], [100]. In theory, the *russell' circumplex model* of emotion allows a *mapping* (diarisation) from continuous signals to emotion labels, which is often refuted [101]. To date, no reliable approaches exist for dynamically modeling continuous emotional values to classes on a large scale.

The question arises whether a special mapping can be derived for 'in-the-wild' environments. This would also help enabling transfer-learning capabilities between the two categories of research and accompanying datasets.

Connected to the previously introduced direction, the continuous annotation signals also have to be *summarised* time-wise on e.g., sentence- or segment-level for the mapping process. Emotions are an intense feeling that is rather short-term and is typically directed at a *source/topic*. We want to explore suitable ways to *time-aggregated emotion annotations* to reference topics. According fusion and aggregation approaches might even rely directly on the five raw annotations, such as in Dynamic Time Warping (DTW) [102] and Deep Canonical Time Warping (DCTW) [103] and might incorporate an explicit model of uncertainty expressed by the (dis-)agreement of annotators. Starting with unsupervised approaches, one might extend the dataset to specific classes in the future. Furthermore, one should explore ways to generally improve the time and cost of intensive continuous annotation. One idea is to learn stable representations of annotators behaviours to create additional artificial annotations based on a reduced number of real annotations.

Since multimodal sentiment analysis 'in-the-wild' often utilises user-generated data and a variety of 'in-the-wild' characteristics, we want to explore those directions in more depth.

With the collected *YouTube metadata* of user engagement (e.g., view count, like and dislike ratio/count, sentimentality of the video comments), we want to investigate if they can be predicted purely relying on (features of) annotated continuous emotions.

The investigation of 'in-the-wild' influences and data imbalance is of further interest. Typically, when developing models for emotion recognition, the aim is to work with a dataset that is as balanced and clean as possible with regard to the people and environment characteristics. This avoids bias in the modelling which leads to worse results when predicting data with an unknown distribution, as well having ethical implications in real-world implementation if not counteracted. When collecting data 'in-the-wild', however, this is sometimes not possible. For this reason, the community needs to explore the effects as well as develop appropriate counter-strategies of, e.g., gender-wise unbalanced training data. Another example is the relevance of *facial-related features* in the prediction of Valence. While in most lab settings it is rather unusual for participants to wear glasses or sunglasses, this is more often the case for 'in-the-wild' data. Previous studies showed that the degree of occlusion of the face is negatively correlated with the performance of face recognition [104] and the *presence of occlusion*, such as sunglasses or masks, degrades the performance of facial expression recognition systems [105]. We have deliberately collected around 30 percent hosts wearing glasses or sunglasses in order to be able to deepen these studies and also estimate how multimodal approaches are of help to overcome those kinds of challenges. A similar directions is to measure the influence of *superimposed banners* which might obscure important video content, such as interaction objects. We want to explore ways to detect, *artificially remove* and replace such occlusion e.g., by utilising Generative Adversarial Networks.

Limitations. As defined by the categories of [68], collecting meaningful in-the-wild data is always a trade-off of naturalness and lack of control. Although we provide natural, minutes-long context to capture the change in emotions towards multiple topics instead of short dialogues, we see some limitation regarding the range of possible context. In this corpus, we wanted to connect emotions to overarching topics enabling emotion-context interaction; hence, only selected material of one domain (car reviews) was considered which naturally limits the range of topic. This has also been done in other large emotional datasets as in, e.g., advertisement in the case of [16]. We see improvements on the other criteria: Primitive descriptors are only common in affective computing but not in multimodal sentiment. We address the gap by providing two dimensions of real continuous emotions and a user-generated content specific one giving advanced flexibility and help to bridge the two communities. The scope of the dataset (e.g., number of hosts, modalities), based on the previous results, does seem to be sufficient to generalise personal-independent affect well and the provided linguistic transcriptions are of high value. Focusing on the user-generated area, we kept a high degree of naturalness, only excluding parts not sensitive to affective recognition (e.g., no face and no voice) while keeping all others, even very noisy sections (e.g., occlusions, selfie-camera, or half faces etc.).

6 Summary and Conclusion

In this paper, we introduced MuSe-CaR – a multimodal sentiment analysis in real-life media dataset. It was collected in user-generated, noisy environments, and consists of around 300 audio-visual and transcript recordings of more than 70 hosts. We described the extensive annotation process in depth covering 11 tiers including dimensional emotions and layers to model the interaction between them and speaker topics and to visual entities. We intentionally selected videos containing novel and challenging in-the-wild characteristics, including dynamic backgrounds and changing shots as well as angles of the face.

From this multimodal corpus of emotional car reviews, we derived three initial tasks: i) MuSe-Wild , where the level of the affective dimensions of Arousal and Valence has to be predicted; ii) MuSe-Topic , where the domain-related conversational topics and three intensity classes of Arousal and Valence have to be predicted; and, iii) MuSe-Trust , where the level of continuous Trustworthiness has to be predicted. These task are publicly available to the research community, representing a testing bed for efforts in automatic analysis of audio-visual behaviour. In addition, we proposed a simple but efficient network DeepTrust using attention-enhanced encoding to tackle the last task, largely outperforming baseline results of MuSe-Trust. We provided exhaustive experiments and showed it has also multi-task prediction capabilities which is both helpful in advancing this novel task as well as the field of continuous affect estimation. Finally, we introduced some of our future research directions and limitations of the dataset. We hope this dataset is another valuable extension for the research community and another cornerstone in mastering multimodal sentiment analysis.

Acknowledgments

The authors would like to thank the annotators and students at the University of Augsburg and Imperial College London who contributed to this project in various ways. This work was supported in part by the the DFG's Reinhart Koselleckproject 442218748 (AUDI0NOMOUS) and in part by the European Union Horizon 2020 research and innovation programme under Grant 856879. Lukas Stappen and Alice Baird are equally contributed to the design of the dataset. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No.115902 (RADAR CNS) and No. 826506 (sustAGE), EPSRC Grant No. 2021037, and the BMW Group.

References

[1] "Cisco visual networking index: Forecast and trends, 2017–2022 white paper," Accessed: Sep. 2019. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11–741490.html

[2] P. Roesler, "Internet video to account for 80% of global traffic by 2019," Accessed: Sep. 2019. [Online]. Available: https://www.webmarketingpros.com/internet-video-to-account-for-80-of/-global-traffic-by-2019/

[3] J. Constine, "Facebook tests video feed to sidestep youtube with friendly discovery," Accessed: Sep. 2019. [Online]. Available: https://techcrunch.com/2015/10/13/facebook-video-feed/

[4] J. Clement, "Most famous social network sites worldwide as of july 2019, ranked by number of active users (in millions)," Accessed: Sep. 2019. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-/ranked-by-number-of-users/

[5] B. Bhardwaj, "Text mining, its utilities, challenges and clustering techniques," *Int. J. Comput. Appl.*, vol. 135, no. 7, pp. 975–8887, 2016.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[7] J. Camacho-Collados and M. T. Pilehvar, "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis," in *Proc. EMNLP Workshop BlackboxNLP: Analyzing Interpreting Neural Netw. NLP*, 2018, pp. 40–46.
[8] X. Chen, Y. Wang, and Q. Liu, "Visual and textual sentiment analysis using deep fusion convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 1557–1561.
[9] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2017, pp. 1103–1114.
[10] M. P. Fortin and B. Chaib-draa, "Multimodal sentiment analysis: A multitask learning approach," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods*, 2019, pp. 368–376.
[11] L. Stappen, V. Karas, N. Cummins, F. Ringeval, K. Scherer, and B. Schuller, "From speech to facial activity: Towards cross-modal sequence-to-sequence attention networks," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2019, pp. 1–6.
[12] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, 2017.
[13] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.
[14] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, 2018. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/12024
[15] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
[16] J. Kossaifi *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1022–1040, Mar. 2021.
[17] L. Stappen *et al.*, "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in *Proc. Int. Multimodal Sentiment Anal. Real-life Media Challenge Workshop, Co-located ACM Int. Conf. Multimedia*, 2020, pp. 35–44.
[18] M. Wöllmer *et al.*, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 46–53, May/Jun. 2013.
[19] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. Int. Conf. Multimodal Interfaces*, 2011, pp. 169–176.
[20] V. Karas and B. W. Schuller, "Deep learning for sentiment analysis: An overview and perspectives," *in Natural Language Processing for Global and Local Business*. Hershey, PA, USA: IGI Global, 2020.
[21] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.2981446.
[22] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, 2018.
[23] D. Kollias *et al.*, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *Int. J. Comput. Vis.*, vol. 127, pp. 1–23, 2019.
[24] E. Douglas-Cowie *et al.*, "The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2007, pp. 488–500.
[25] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Proc. IEEE Int. Conf. Workshops Automat. Face Gesture Recognit.*, 2013, pp. 1–8.
[26] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "Afew-va database for valence and arousal estimation in-the-wild," *Image Vis. Comput.*, vol. 65, pp. 23–36, 2017.
[27] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2008, pp. 865–868.
[28] C. Busso *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, 2008, Art. no. 335.
[29] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan.–Mar. 2011.
[30] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The belfast induced natural emotion database," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 32–41, Jan.–Mar. 2012.
[31] M. K. Hasan *et al.*, "Ur-funny: A multimodal language dataset for understanding humor," in *Proc. Conf. Empirical Methods Nat. Lang. Process. Int. Joint Conf. Nat. Lang. Process.*, 2019, pp. 2046–2056.
[32] A. B. Zadeh, Y. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, "Moseas: A multimodal language dataset for spanish, portuguese, german and french," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2020, pp. 1801–1812.
[33] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[34] E. Marrese-Taylor, C. Rodriguez-Opazo, J. A. Balazs, S. Gould, and Y. Matsuo, "A multi-modal approach to fine-grained opinion mining on video reviews," 2020, *arXiv:2005.13362*.
[35] W. Yu *et al.*, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.
[36] D. Cevher, S. Zepf, and R. Klinger, "Towards multimodal emotion recognition in german speech events in cars using transfer learning," *in Proc. Conf. Nat. Lang. Process.*, 2019, pp. 79–90.
[37] E. Marrese-Taylor, J. Balazs, and Y. Matsuo, "Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN," in *Proc. Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, Sep. 2017, pp. 102–111.
[38] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 973–982.
[39] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Proc. Mining Text Data*, 2012, pp. 415–463.
[40] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—A review of research topics, venues, and top cited papers," *Comput. Sci. Rev.*, vol. 27, pp. 16–32, 2018.
[41] V. P. Rosas, R. Mihalcea, and L.-P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 38–45, May/Jun. 2013.
[42] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 223–232.
[43] A. Garcia, S. Essid, F. d'Alché Buc, and C. Clavel, "A multimodal movie review corpus for fine-grained opinion mining," 2019, *arXiv:1902.10102*.
[44] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency, "Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach," in *Proc. Int. Conf. Multimodal Interact.*, 2014, pp. 50–57.
[45] J. A. Russell, "A circumplex model of affect," *J. Pers. Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.
[46] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010.
[47] S. M. Mohammad, "Sentiment analysis: Detecting valence, emotions, and other affectual states from text," in *Proc. Emotion Meas*, 2016, pp. 201–237.
[48] D. Preo ţiuc-Pietro *et al.*, "Modelling valence and arousal in facebook posts," in *Proc. Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2016, pp. 9–15.
[49] D. Kollias and S. Zafeiriou, "Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–15.
[50] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds Mach.*, vol. 30, pp. 681–694, 2020.
[51] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 353–355.

[52] A. Luneski, E. Konstantinidis, and P. Bamidis, “Affective medicine: A review of affective computing efforts in medical informatics,” *Methods Inf. Med.*, vol. 49, no. 3, pp. 207–218, 2010.

[53] A. Baird and B. Schuller, “Considerations for a more ethical approach to data in AI: On data representation and infrastructure,” *Front. Big Data*, vol. 3, 2020, Art. no. 25.

[54] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, “Elan: A professional framework for multimodality research,” in *Proc. Int. Conf. Lang. Resour. Eval.*, 2006, pp. 1556–1559.

[55] L. C. De Silva, T. Miyasato, and R. Nakatsu, “Facial emotion recognition using multi-modal information,” in *Proc. Int. Conf. Inf., Commun. Signal Process. Theme: Trends Inf. Syst. Eng. Wirel. Multimedia Commun.*, 1997, pp. 397–401.

[56] L. F. Barrett, B. Mesquita, and M. Gendron, “Context in emotion perception,” *Curr. Directions Psychol. Sci.*, vol. 20, no. 5, pp. 286–290, 2011.

[57] J. M. Girard and A. G. Wright, “Darma: Software for dual axis rating and media annotation,” *Behav. Res. Methods*, vol. 50, no. 3, pp. 902–909, 2018.

[58] C. Schwemmer and S. Ziewiecki, “Social media sellout: The increasing role of product promotion on youtube,” *Social Media+ Soc.*, vol. 4, no. 3, 2018.

[59] S. T. Moturu and H. Liu, “Quantifying the trustworthiness of social media content,” *Distrib. Parallel Databases*, vol. 29, no. 3, pp. 239–260, 2011.

[60] H. Horsburgh, “Trust and social objectives,” *Ethics*, vol. 72, no. 1, pp. 28–40, 1961.

[61] J. C. Cox, R. Kerschbamer, and D. Neururer, “What is trustworthiness and what drives it?,” *Games Econ. Behav.*, vol. 98, pp. 197–218, 2016.

[62] J. A. Colquitt, B. A. Scott, and J. A. LePine, “Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance,” *J. Appl. Psychol.*, vol. 92, no. 4, 2007, Art. no. 909.

[63] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic, “Robust correlated and individual component analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1665–1678, Aug. 2016.

[64] B. W. Schuller, *Intelligent Audio Analysis*. Berlin, Germany: Springer, 2013.

[65] M. Grimm and K. Kroschel, “Evaluation of natural emotions using self assessment manikins,” in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, 2005, pp. 381–385.

[66] S. Hantke, E. Marchi, and B. Schuller, “Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification,” in *Proc. Int. Conf. Lang. Resour. Eval.*, 2016, pp. 2156–2161.

[67] F. Ringeval *et al.*, “Avec 2017: Real-life depression, and affect recognition workshop and challenge,” in *Proc. Annu. Workshop Audio/Visual Emotion Challenge*, 2017, pp. 3–9.

[68] E. Douglas-Cowie *et al.*, “Multimodal databases of everyday emotion: Facing up to complexity,” in *Proc. Eur. Conf. Speech Commun. Technol.*, 2005, pp. 813–816.

[69] L. Devillers, L. Vidrascu, and L. Lamel, “Challenges in real-life emotion annotation and machine learning based detection,” *Neural Netw.*, vol. 18, no. 4, pp. 407–422, 2005.

[70] A. Słucki, T. Trzciński, A. Bielski, and P. Cyrta, “Extracting textual overlays from social media videos using neural networks,” in *Proc. Int. Conf. Comput. Vis. Graph.*, 2018, pp. 287–299.

[71] S. Tian, Y. Pan, C. Huang, S. Lu, K. Yu, and C. Lim Tan, “Text flow: A unified text detection system in natural scene images,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4651–4659.

[72] H. Yang, C. Wang, C. Bartz, and C. Meinel, “Scenetextreg: A realtime video OCR system,” in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 698–700.

[73] H. Su, J. Deng, and L. Fei-Fei, “Crowdsourcing annotations for visual object detection,” in *Proc. Workshops AAAI Conf. Artif. Intell.*, 2012, pp. 40–46.

[74] K. Konyushkova, J. Uijlings, C. H. Lampert, and V. Ferrari, “Learning intelligent dialogs for bounding box annotation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9175–9184.

[75] L. Stappen, X. Du, V. Karas, S. Müller, and B. W. Schuller, “Gocard–generic, optical car part recognition and detection: Collection, insights, and applications,” 2020, *arXiv0: 2006.08521*.

[76] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[77] T. Afouras, J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition,” *CoRR*, abs/1809.02108, 2018. [Online]. Available: http://arxiv.org/abs/1809.02108

[78] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3730–3738.

[79] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” *in Proc. Brit. Mach. Vis. Assoc.*, 2015, pp. 41.1–41.12. [Online]. Available: http://www.bmva.org/bmvc/2015/papers/paper041/index.html

[80] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “Openface: An open source facial behavior analysis toolkit,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[81] M. Valstar *et al.*, “Avec 2013: The continuous audio/visual emotion and depression recognition challenge,” in *Proc. ACM Int. Workshop Audio/Visual Emotion Challenge*, 2013, pp. 3–10.

[82] V. Pandit and B. Schuller, “On many-to-many mapping between concordance correlation coefficient and mean square error,” 2019, *arXiv:1902.05180*.

[83] I. Lawrence and K. Lin, “A concordance correlation coefficient to evaluate reproducibility,” *Biometrics*, vol. 45, pp. 255–268, 1989.

[84] L. Stappen, B. W. Schuller, I. Lefter, E. Cambria, and I. Kompatsiaris, “Summary of muSe 2020: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media,” in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 4769–4770.

[85] F. Eyben *et al.*, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2015.

[86] S. Hershey *et al.*, “CNN architectures for large-scale audio classification,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 131–135.

[87] J. F. Gemmeke *et al.*, “Audio set: An ontology and human-labeled dataset for audio events,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2017, pp. 776–780.

[88] T. Baltrušaitis, P. Robinson, and L. Morency, “OpenFace: An open source facial behavior analysis toolkit,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.

[89] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Trans. Assoc. Comput. Linguistics*, vol. 5, pp. 135–146, 2017.

[90] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, 2019, Art. no. 9.

[91] 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 *Proc. Int. Multimodal Sentiment Anal. Real-life Media Challenge Workshop*, 2020, pp. 27–34.

[92] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[93] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, “Efficient modeling of long temporal contexts for continuous emotion recognition,” in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2019, pp. 185–191.

[94] H.-J. Yang, G.-S. Lee, J.-H. Kim, and S.-H. Kim, “Multimodal fusion with attention mechanism for trustworthiness prediction in car advertisements,” in *Proc. 9th Int. Conf. Smart Media Appl.* 2020. [Online]. Available: https://sma-2020.github.io/accepted_papers/

[95] A. Zadeh, P. P. Liang, L.-P. Morency, S. Poria, E. Cambria, and S. Scherer, “Proceedings of grand challenge and workshop on human multimodal language (challenge-hml),” in *Proc. Grand Challenge Workshop Hum. Multimodal Lang.*, 2018. [Online]. Available: https://aclanthology.org/W18-3300/

[96] D. Kollias, A. Schulc, E. Hajiyev, and S. Zafeiriou, “Analysing affective behavior in the first abaw 2020 competition,” 2020, *arXiv:2001.11409*.

[97] B. W. Schuller *et al.*, “The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats,” in *Proc. Interspeech*, 2018, pp. 122–126.

[98] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González, “Gated multimodal networks,” *Neural Comput. Appl.*, vol. 32, pp. 10209–10228, 2020.

[99] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, “Exploring hate speech detection in multimodal publications,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1470–1478.

[100] X. Qiu, Z. Feng, X. Yang, and J. Tian, "Multimodal fusion of speech and gesture recognition based on deep learning," in *Proc. J. Phys.: Conf. Series*, 2020, pp. 458–466.

[101] S. Hamann, "Mapping discrete and dimensional emotions onto the brain: controversies and consensus," *Trends Cogn. Sci.*, vol. 16, no. 9, pp. 458–466, 2012.

[102] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Hoboken, NJ, USA: Prentice-Hall, 1993.

[103] G. Trigeorgis, M. A. Nicolaou, S. Zafeiriou, and B. W. Schuller, "Deep canonical time warping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5110–5118.

[104] R. R. Atallah, A. Kamsin, M. A. Ismail, S. A. Abdelrahman, and S. Zerdoumi, "Face recognition and age estimation implications of changes in facial features: A critical review study," *IEEE Access*, vol. 6, pp. 28 290–28 304, 2018.

[105] M. Sajjad, A. Shah, Z. Jan, S. I. Shah, S. W. Baik, and I. Mehmood, "Facial appearance and texture feature-based robust facial expression recognition framework for sentiment knowledge discovery," *Cluster Comput.*, vol. 21, no. 1, pp. 549–567, 2018.

Lukas Stappen (Member, IEEE) received the Master of Science degree in data science with distinction from King's College London in 2017. He is currently working toward the PhD degree with the Chair for Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, and a PhD fellow with the BMW Group. He then joined the group for Machine Learning in Health Informatics. His research interests include affective computing, multimodal sentiment analysis, and multimodal or cross-modal representation learning with a core focus on 'in-the-wild' environments.

Alice Baird (Member, IEEE) received the MFA degree in sound art from Columbia University's Computer Music Center. She is currently a PhD fellow with the ZD.B, supervised by Prof. Björn Schuller with the Chair of Embedded Intelligence for Healthcare and Wellbeing, University of Augsburg, Germany. Her research is focused on intelligent audio analysis in the domain of speech and general audio, and her research interests include health informatics, affective computing, computational paralinguistics, and speech pathology.

Lea Schumann received the BSc degree in computer science from the University of Augsburg, Germany, in 2018. She is currently working toward the MSc degree in computer science with a strong focus on deep learning. Her research interests include computer vision, affective computing, and natural language processing.

Björn Schuller (Fellow, IEEE) received the Diploma, doctoral degree, habilitation, and adjunct teaching professor degrees in machine intelligence and signal processing in EE/IT from TUM in Munich, Germany. He is currently a full professor of artificial intelligence and the Head of GLAM, Imperial College London, U.K., full professor and Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, and a permanent visiting professor with HIT, China, amongst other professorships and affiliations. He was a full professor with the University of Passau, Germany, a researcher with Joanneum Research, Graz, Austria, and the CNRS-LIMSI with Orsay, France. He is a fellow of Golden Core Awardee of the IEEE Computer Society, BCS, ISCA, President-Emeritus of the AAAC, and a senior member of the ACM. He has authored or coauthored more than 1000 publications, more than 35000 citations, and 85 h-index. He is the field chief editor of *Frontiers in Digital Health* and was the editor in chief of the *IEEE Transactions on Affective Computing* amongst manifold further commitments and service to the community. He was the recipient of more than 30 awards include having been honoured as one of 40 extraordinary scientists under the age of 40 by the WEF in 2015.