MuSe 2021 Challenge: Multimodal Emotion, Sentiment, Physiological-Emotion, and Stress Detection

Lukas Stappen
EIHW, University of Augsburg
Augsburg, Germany

Eva-Maria Meßner
University of Ulm
Ulm, Germany

Erik Cambria
Nanyang Technological University
Singapore

Guoying Zhao
University of Oulu
Oulu, Finland

Björn W. Schuller
GLAM, Imperial College London
London, United Kingdom

ABSTRACT

The 2nd **Mu**ltimodal **Se**ntiment Analysis (MuSe) 2021 Challenge-based Workshop is held in conjunction with ACM Multimedia'21. Two datasets are provided as part of the challenge. Firstly, the MuSe-CaR dataset, which focuses on user-generated, emotional vehicle reviews from YouTube, and secondly, the novel Ulm-Trier Social Stress (Ulm-TSST) dataset, which shows people in stressful circumstances. Participants are faced with four sub-challenges: predicting arousal and valence in a time- and value-continuous manner on a) MuSe-CaR (MuSe-Wilder) and b) Ulm-TSST (MuSe-Stress); c) predicting unsupervised created emotion classes on MuSe-CaR (MuSe-Sent); d) predicting a fusion of human-annotated arousal and measured galvanic skin response also as a continuous target on Ulm-TSST (MuSe-Physio). In this summary, we describe the motivation, the sub-challenges, the challenge conditions, the participation, and the most successful approaches.

1 INTRODUCTION AND MOTIVATION

This year's **Mu**ltimodal **Se**ntiment Analysis in Real-life Media Challenge (MuSe) and workshop has been organised in conjunction with the 29th ACM International Conference on Multimedia was held as a hybrid conference in Chengdu, China. The MuSe 2021 is the second competition event aimed at comparison of multimedia processing from the automatic transcribed spoken word, audio, and visual modalities and machine learning methods to predict value and time continuous arousal, physiological-arousal, valence, and discrete classes derived from the continuous annotations. In addition, the MuSe series [6, 8] bridges the text-centred sentiment analysis community [2, 4] and signal-focused affective computing community [7].

In the 2021 edition, we called for four distinct sub-challenges, each focusing on specific aspects of a united viewpoint: predicting the a) MuSe-Wilder : level of emotional, time-continuous dimensions (arousal, valence) in user-generated audio-visual recordings; b) MuSe-Sent : time and value summarised advanced emotion classes derived from arousal and valence annotations by using the MuSe-Toolbox [9] on segment-level of the same recordings; c) MuSe-Stress : level of emotional arousal and valence in a time-continuous manner capturing people in stressful interview dispositions; and d) MuSe-Physio : level of physiological arousal derived from human annotations fused with electrodermal activity signals.

As benchmark datasets, last year's dataset [6], MuSe-CaR [7] with 40 hours of video is re-used for the sub-challenges a) and b) featuring novel, substantially improved methods for gold-standard creation. Furthermore, the Ulm-TSST corpus is introduced, which displays people faced with the Ulm-Trier Social Stress Test. Both datasets include automatically transcribed spoken language to which the audio-video features were aligned. Besides the data and the code for the baseline models, a broad selection of language, audio, and visual features were provided [5]. The goal of the challenge is to raise the bar for prediction outcomes, modelling approaches, and handling of naturalistic environments of these research areas. The novel methods developed by the participants are compared under strictly the same conditions. While the MuSe-CaR dataset is fully available to the academic research community outside the challenge, the Ulm-TSST will only be shared to selected research parties until the current research projects are completed. The key statistics can be found in Table 1.

Table 1: Registrations per sub-challenges relative to total; unique downloads per data package; distribution of test submissions; (combined arousal and valence) baseline result and best participants result on the test set.

	MuSe-Wilder	MuSe-Sent	MuSe-Stress	MuSe-Physio
Registrations [%]	45.7	69.6	57.6	42.4
Downloads	91	135	73	34
Test submissions [%]	4.2	50.3	34.0	11.5
Baseline [5]	.4616	32.82	.5088	.4908
Best challenge result	.5413	40.33	.5384	.5781

2 CHALLENGE PROCEDURE

To gain access to the data, a team lead was required to hold an academic position and sign an end-user licence agreement. After reviewing the paperwork, they received a link to the research data platform zenodo.org [1]. Here, team members could download the raw and sub-challenge specific data packages. The raw data contained the anonymised raw video and audio files, as well as the metadata. The task-specific data packages contained the feature sets and labels as used in the baselines [6]. Hereby, the ground truth labels of the test set were masked and could only be utilised to prepare the participants' test set predictions. The MuSe-CaR raw data recorded 304 downloads and Ulm-TSST raw 148 (detailed see Table 1).

Furthermore, the baseline code was also made publicly available on GitHub [2] and contains a detailed description and the weights

[1] MuSe-CaR raw: https://zenodo.org/record/4651164; MuSe-Wilder: https://zenodo.org/record/4652376, MuSe-Sent: https://zenodo.org/record/4654371; Ulm-TSST raw: https://doi.org/10.5281/zenodo.4767117, MuSe-Stress: https://doi.org/10.5281/zenodo.4767114, MuSe-Physio: https://doi.org/10.5281/zenodo.4765992.
[2] https://github.com/lstappen/MuSe2021

of the most successful networks. With the combination of code and data, participants were able to repeat the baseline experiments. Upon finding their most appropriate method, participants could submit their test set prediction to the portal[3] one month before the paper deadline. The MuSe data chairs scored the submission, so that the results were obtained by the submitter within one day. Each team had up to five attempts per sub-challenge. The codebases of the best performing teams underwent a sanity check.

3 PARTICIPATION

The call for participation attracted registrations of 61 teams from 16 countries and 47 academic institutions. Around half of the teams submitted their test results, resulting in a total of 189 test set submissions. Compared to the 2020 MuSe challenge, the participation has almost tripled [6, 8]. After closing the test scoring, teams were either informed that they had finished in the top 3 or were given their exact rank if worse. Afterwards, (top) participants had to lay out their developed approach and submit a paper of up to 8 pages (plus references). In total, 14 valid paper submissions were received. The submitted papers were double-blind reviewed by at least three members of the programme committee for scientific quality, novelty, and technical correctness. Papers were accepted if all reviewers agreed to accept.

4 CHALLENGE OUTCOME

A brief overview of the results and a summary from the submitted papers are given. Although the baseline level was tough this year, all baseline results were exceeded. The relative improvement over the baseline result is given in per cent on the test set. For MuSe-Wilder, the best models improved the Concordance Correlation Coefficient (CCC) on arousal by 23 % (.3386 to .4177) and on valence roughly by 11% (.5974 to .6649) compared to the baseline. The only classification task, MuSe-Sent, saw improved F1 score results by 3 % for arousal (35.12 % to 36.14 %) and by 35 % for valence (32.91 % to 44.51 %). The considerable improvements on the MuSe-CaR sub-challenges can largely be attributed to the increased exploitation of data-driven, context-dependent feature extraction based on transformer networks. For example, DeBERTa [3] was extracted from the transcripts and wav2vec [1] from the audio track. In addition, attention-based networks were increasingly successful.

On the Ulm-TSST dataset, the best results of MuSe-Stress on arousal were only slightly above the baseline models, with 1 % improvement (.4562 to .4609 CCC), while valence was outperformed by 18 % (.5614 to .6648 CCC). On the experimental task MuSe-Physio, the best result achieved .5781 CCC up from .4908 CCC. As before, attention-based networks such as Transformer Encoder with Multi-modal Multi-head Attention and Transformer-based feature extractions were beneficial in improving the baseline. In addition, the biosignals (e. g., heart rate, and respiration), which were also made available as features, led to performance improvements. This illustrates the close link between human-annotated and physically measured biosignals in arousal-intensive prediction tasks.

[3]https://www.muse-challenge.org

5 WORKSHOP ORGANISATION

MuSe takes place as a full-day, virtual workshop. The program features a short introduction and closing note provided by the workshop organisers, two invited keynote speeches, as well as oral presentations of the accepted papers.

We appreciate the reviewers' efforts and would like to thank the data chairs Alice Baird (University of Augsburg, DE) and Lukas Christ (University of Augsburg, DE) as well as the program committee for their valuable support: Elisabeth André (University of Augsburg, DE), Paul Buitelaar (National University of Ireland, IE), Carlos Busso (The University of Texas at Dallas, US), Oana Cocarascu (Imperial College London, UK), Dipankar Das (Jadavpur University, IN), Alexander Gelbukh (Instituto Politécnico Nacional, MX), Gil Keren (Facebook, US), Iftekhar Naim (Google, US), Preslav Nakov (UC Berkeley, US), Symeon Papadopoulos (ITI, GR), Ioannis Patras (Queen Mary University of London, UK), Peter Robinson (University of Cambridge, UK), Mohammad Soleymani (USC, US), and Alessandro Vinciarelli (University of Glasgow, UK), and Yiqun Yao (University of Michigan, US).

REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems* 33 (2020).

[2] Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and RBV Subramanyam. 2017. Benchmarking multimodal sentiment analysis. In *International Conference on Computational Linguistics and Intelligent Text Processing*. Springer, 166–179.

[3] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations (ICLR)*.

[4] Lukas Stappen, Alice Baird, Erik Cambria, and Björn W Schuller. 2021. Sentiment Analysis and Topic Recognition in Video Transcriptions. *IEEE Intelligent Systems* 36, 2 (2021), 88–95.

[5] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Messner, Erik Cambria, Guoying Zhao, and Björn W Schuller. 2021. The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress. In *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge*. ACM.

[6] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Bjoern W. Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. 2020. MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-Target Engagement and Trustworthiness Detection in Real-Life Media. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*. ACM, 35–44.

[7] Lukas Stappen, Alice Baird, Lea Schumann, and Björn Schuller. 2021. The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements. *IEEE Transactions on Affective Computing* 01 (06 2021), 1–16.

[8] Lukas Stappen, Björn Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. 2020. Summary of MuSe 2020: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4769–4770.

[9] Lukas Stappen, Lea Schumann, Benjamin Sertolli, Alice Baird, Benjamin Weigel, Erik Cambria, and Björn W Schuller. 2021. MuSe-Toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox. In *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge*. ACM.