Summary of MuSe 2020: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media

Lukas Stappen
University of Augsburg
Augsburg, Germany

Björn W. Schuller
Imperial College London
London, United Kingdom

Iulia Lefter
Delft University of Technology
Delft, Netherlands

Erik Cambria
Nanyang Technological University
Singapore

Ioannis Kompatsiaris
CERTH - ITI
Thermi-Thessaloniki, Greece

ABSTRACT

The first **Mu**ltimodal **Se**ntiment Analysis in Real-life Media (MuSe) 2020 was a Challenge-based Workshop held in conjunction with ACM Multimedia'20. It addresses three distinct 'in-the-wild' Sub-challenges: *sentiment/ emotion recognition* (MuSe-Wild), *emotion-target engagement* (MuSe-Target) and *trustworthiness detection* (MuSe-Trust). A large multimedia dataset *MuSe-CaR* was used, which was specifically designed with the intention of improving machine understanding approaches of how sentiment (i. e., emotion) is linked to a topic in emotional, user-generated reviews. In this summary, we describe the motivation, the first of its kind 'in-the-wild' database, the challenge conditions, the participation, as well as give an overview of utilised state-of-the-art techniques.

CCS CONCEPTS

• **Information systems → Multimedia and multimodal retrieval**.

KEYWORDS

Multimodal Sentiment Analysis; Affective Computing; User-Generated Data; Multimodal Fusion

1 INTRODUCTION AND MOTIVATION

The **Mu**ltimodal **Se**ntiment Analysis in Real-life Media (MuSe) 2020 Challenge and Workshop aimed at giving researchers the opportunity to apply novel methods for the fusion of the audio-visual and language modalities under strictly the same conditions, and bringing together communities from differing disciplines; such as the sentiment analysis community (symbol-based), and the audio-visual emotion recognition community (signal-based).

The field of Sentiment Mining specialising in Natural Language Processing methods for symbolic information analysis leverages text modality and focuses on the prediction only of discrete sentiment label categories [4]. Lately, more approaches considered additional modalities [1]. Coming from the origin of intelligent signal processing, the emotion recognition community, belonging to the field of Affective Computing, mostly focuses on one, or both of the audio and vision modalities, in order to predict the continuous-valued valence and arousal dimensions of emotion (circumplex model of affect), while often disregarding the potential contribution of textual information [2, 6]. Recently, approaches have become more comprehensive, and explicitly comprise all three modalities [3, 7].

The main underlying motivation is the need to advance the fields of multimodal sentiment and emotion recognition for multimedia retrieval to a point where content and behaviour expressed during interactions can be reliably sensed and captured in real-world conditions. We introduced the Multimodal Sentiment Analysis in Car Reviews dataset *MuSe-CaR* consisting of almost 40 hours of video in this year's challenge. Utilising the MuSe-CaR dataset we called for three distinct Sub-challenges: MuSe-Wild , which focuses on continuous emotion (valence/sentiment and arousal) prediction; MuSe-Topic , in which participants recognise 10 domain-specific topics as the target of 3-class (low, medium, high) emotions; and MuSe-Trust , in which the novel aspect of trustworthiness is to be predicted. Data sets used in similar previous challenges displaying natural behaviour [2, 6] were recorded under rather strictly controlled conditions, while MuSe-CaR contains user-generated multimedia influenced by new 'in-the-wild' characteristics, e. g., changing camera settings (zoom, shots), varying backgrounds, and ambient noises.

Ideally, the participants of MuSe strove towards the development of unified approaches, which are applicable to any of the tasks offered through the challenge, in this way, understanding the degree to which fusion is possible, as well as offering advancements for sentiment and emotion recognition systems that deal with fully naturalistic (in-the-wild) behaviour from large volumes of data.

To save participants the time-consuming extraction of features, which commonly takes days or weeks when dealing with large amounts of data, and shifting the focus towards modelling, a broad selection of extracted language, audio, and visual features as well as baseline models for a first evaluation of the strengths of the core modalities were provided [5].

2 CHALLENGE CONDITIONS AND SELECTION PROCESS

For participation and access to the data of the MuSe Challenge, we required the participants to sign an end-user license agreement. The

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5
https://doi.org/10.1145/3394171.3421901

data package itself came with the metadata, raw video files, and 14 model-ready audio, visual, and linguistic feature sets pre-computed for each Sub-challenge. The code repository for reproducing any of the baselines [5] was made publicly available for attendees and the interested reader[1]. After downloading the data, they could start to develop their approaches using the training and development partitions. The ground truth labels of the test set were masked, so that this data partition could only be used to prepare the predictions for scoring. Upon finding their most appropriate method for any of the three Sub-challenges, participants could submit their test set predictions online and received the results within one working day as an email from the MuSe data chairs. Each team had up to five submission attempts per Sub-challenge.

To present their successful approach, a paper with up to 8 pages had to be written and submitted. All paper submissions were reviewed double blindly by at least three members of the program committee with respect to scientific quality, novelty, technical quality, and the suitability to the workshop's topic. Submissions were accepted if all reviewers agreed to accept.

3 PARTICIPANTS AND OUTCOME

The call for participation attracted registrations of 21 teams from 11 countries and 16 academic institutions. The codebases of the best performing teams underwent a sanity check. Although not all participants submitted their approaches as papers or got accepted by the reviewers, the following is a brief, overarching overview of the results on the test set and the first findings. For predicting the time-continuous emotional dimensions (MuSe-Wild), the best models improved the CCC on arousal roughly by 0.19 (0.2834 to 0.4726) and on valence roughly by 0.36 (0.2431 to 0.5996) compared to the baseline. Almost all neural network architectures were based on (self-)attention enhanced Recurrent Neural Networks. In line with previous research, multimodal fusion approaches surpassed uni-modal models, with textual feature sets (e. g., Bert) being most predictive for sentiment/ valence, while audio feature sets (e. g., VGGish) proved to be most valuable in predicting arousal. In discordance with previous findings of the well-established Audio-Visual Emotion Challenge, the vision feature sets achieved lower results than the audio feature sets. A plausible reason for this could be that the MuSe data set has the most challenging 'in-the-wild' characteristics in the visual modality (non-frontal reviewers etc.). Regardless of modality, participants using the CCC loss achieved higher results than those using an L1 or MSE loss.

No submissions were received outperforming the highly competitive baseline model results (CCC: 0.4128) for the task of predicting the time-continuous level of trustworthiness (MuSe-Trust).

The best results of MuSe-Target were only slightly above the baseline models. The combined score (valence and arousal) improved compared to the baseline from 38.81 to 40.33 (in this setting, arousal improved from 42.67 to 45.16, while valence worsen from 40.12 to 39.75 percent), while the score predicting the topic increased marginally from 76.78 to 77.08 percent. We assume that the formed emotion classes could profit from a more advanced emotion diarisation.

The results of the challenge and contributions included in the proceedings clearly show that analysing user-generated multimedia regarding sentiment/ emotion and context is far from solved, requiring further efforts in the future.

4 WORKSHOP ORGANISATION

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

We appreciate the reviewers' efforts and would like to thank the members of the data chair Alice Baird (University of Augsburg, DE) and Georgios Rizos (Imperial College London, UK) 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), Alessandro Vinciarelli (University of Glasgow, UK), Rui Xia (Nanjing University of Science and Technology, CN), and Zixing Zhang (Huawei Technologies Research, UK).

REFERENCES

[1] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2359–2369.

[2] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. Avec 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 3–9.

[3] Björn W Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, et al. 2020. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. *Proceedings INTERSPEECH. Shanghai, China: ISCA* (2020).

[4] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of Multimodal Sentiment Analysis. *Image and Vision Computing* 65 (2017), 3–14.

[5] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Björn 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 *1st International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, co-located with the 28th ACM International Conference on Multimedia (ACM MM)*. ACM.

[6] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*. ACM, 3–10.

[7] Amir Zadeh, Louis-Philippe Morency, Paul Pu Liang, Soujanya Poria, Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML). In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, Amir Zadeh, Louis-Philippe Morency, Paul Pu Liang, and Soujanya Poria (Eds.). ACL, Seattle, USA.

[1] https://github.com/lstappen/MuSe2020