Table 2: Reporting Arousal, Valence and Combined ($0.5 \cdot Arousal + 0.5 \cdot Valence$) for MuSe-Wild and Trustworthiness for MuSe-Trust, both using concordance correlation coefficient (CCC). As feature sets FastText (FT), eGeMAPS (Ge), DeepSpectrum (DS), GoCaR (Go), VGGFace (VG), and Xception (X) and all visual (aV) are fed into the models. Furthermore, the raw audio signal (RA) is used in End2You, and low-level descriptors (LDD) are utilised for MuSe-Trust in order to predict trustworthiness. All utilised features of MuSe-Wild and MuSe-Trust are aligned to the label timestamps by imputing missing values or repeating the word embeddings for FastText.

System	Features	Valence devel / test	Arousal devel / test	Combined devel / test	Trustworthiness devel / test
LSTM + Self-ATT	LLD	.0711 / .0349	**.3078 / .2834**	.1894 / .1592	.2560 / .1343
LSTM + Self-ATT	DS	.0165 / .0024	.1585 / .1723	.0875 / .0874	.2019 / .1701
LSTM + Self-ATT	Ge	.0435 / -.0097	.1090 / .0827	.0762 / .0365	.1576 / .1385
LSTM + Self-ATT	FT	.1273 / .1816	.0959 / .1074	.1116 / .1445	.2278 / .2549
LSTM + Self-ATT	Ge + FT	.0520 / .0361	.1375 / .1018	.0947 / .0690	.2296 / .2054
LSTM + Self-ATT	X	.0499 / .0426	.0776 / .0683	.0638 / .0555	.1178 / .1664
LSTM + Self-ATT	aV	.0098 / .0272	.1598 / .1227	.0848 / .0749	.1167 / .1378
LSTM + Self-ATT	Ge + FT + V	.0393 / .0654	.1809 / .0865	.1101 / .0760	.1245 / .1695
End2You	FT + VG + RA	**.1506 / .2431**	.2587 / .2706	**.2047 / .2568**	**.3198 / .4128**
End2You-Multitask	FT + VG + RA	–	–	–	.3264 / .4119

Table 3: MuSe-Topic: Reporting Unweighted Average Recall (UAR), F1, and Combined ($0.66 \cdot F1 + 0.34 \cdot UAR$) for the topic predictions. As feature sets FastText (FT), Raw Text (RT), eGeMAPS (eG), DeepSpectrum (DS), VGGFace (VG), Xception (X), OpenPose (OP), GoCaR (Go), Facial Action Units (AU) and all visual features (aV) are used. Two types of alignment are used to a) align to eGeMAPS (GA), and b) aggregate on FastText word features (FA).

System	Features	Alig.	F1 devel / test	UAR devel / test	Combined devel / test
LSTM + Self-ATT	DS	GA	19.85 / 34.60	12.95 / 35.00	17.50 / 34.74
LSTM + Self-ATT	eG	GA	19.02 / 34.44	12.34 / 33.94	16.75 / 34.27
LSTM + Self-ATT	FT	GA	24.62 / 36.19	15.25 / 36.22	21.44 / 36.20
LSTM + Self-ATT	eG + FT	GA	20.38 / 35.32	13.13 / 34.87	17.92 / 35.16
LSTM + Self-ATT	X	GA	26.06 / 36.83	20.42 / 36.61	24.14 / 36.75
LSTM + Self-ATT	aV	GA	27.42 / 34.92	21.57 / 34.41	25.43 / 34.75
LSTM + Self-ATT	eG + FT + V	GA	27.58 / 37.14	20.08 / 37.14	25.03 / 37.14
Fine-tuned Albert	RT	–	**71.69 / 76.59**	**69.56 / 77.18**	**70.96 / 76.79**
MMT	FT + eG + X	–	48.24 / 53.18	41.49 / 50.44	44.86 / 51.81
MMT	FT + eG + X	FA	49.21 / 52.06	40.78 / 49.68	46.35 / 51.25
MMT	FT + eG + VG	–	44.72 / 50.71	35.60 / 45.19	41.62 / 48.84
MMT	FT + eG + Go	–	44.72 / 53.73	38.14 / 49.85	42.48 / 52.41
MMT	FT + eG + AU	–	46.22 / 54.52	40.66 / 49.99	44.33 / 52.98
MMT	FT + Go + OP	–	45.17 / 52.30	37.82 / 48.61	42.67 / 51.05
MMT	FT + DS + Go	–	44.42 / 52.06	36.77 / 49.59	41.82 / 51.22

achieve results higher than chance level (33 %) on test e. g., the fine-tuned Albert. Overall, the **Multimodal Transformer achieved with** 38.81% **(combined valence and arousal) the best results utilising FastText, eGeMAPS, and Xception**. The same configuration is also **most successful in predicting valence (**40.12%**) on test**. The utilised SVMs, chosen due to their scalability on high dimensional data, showed results comparable to most state-of-the-art approaches. In particular, for the prediction of arousal, the **VGGFace features result is the best Combined F1 and UAR of** 42.67 **on test**. These SVM results lead us to assume that this task may benefit from a more traditional feature-level analysis. The confusion matrix for all tasks are depicted in Figure 2.

6.3 MuSe-Trust

The results for the prediction of trustworthiness are depicted in Table 2. Similar to MuSe-Wild, the end-to-end baseline system using FastText, VGGFace, and raw audio signals gave the best results with .4128 CCC on test. The results may improve if the valence and arousal predicted signals are incorporated during training. This can be accomplished in three ways: i) the model from MuSe-Wild is utilised to predict arousal and valence on MuSe-Trust; ii) the arousal and valence models can be retrained on MuSe-Trust (we provide train and devel labels); or iii) all three are predicted in a multitask-fashion (one model, 3 outputs) on train and devel, and only trustworthiness is predicted on test. We decided for option (iii). Adding these signals to the **end-to-end baseline system**, the predictive power of the model is similar to the previous one with CCC .3264 on the development set and .4119 on the test set.

7 CONCLUSIONS

In this paper, we introduced MuSe 2020 – the first Multimodal Sentiment Analysis in real media assessment challenge. MuSe 2020 utilises the MuSe-CaR multimodal corpus of emotional car reviews and comprises three Sub-challenges: i) MuSe-Wild, where the level of the affective dimensions of valence (corresponding to sentiment) and arousal has to be predicted from a ca. 35 hour data subset; ii) MuSe-Topic, where the domain-related conversational topic (10

Table 4: MuSe-Topic: Reporting Unweighted Average Recall (UAR), F1, and Combined ($0.66 \cdot F1 + 0.34 \cdot UAR$) for the 3-class valence and arousal predictions and the combined (mean) of valence and arousal. As feature sets FastText (FT), Raw Text (RT), eGeMAPS (eG), DeepSpectrum (DS), VGGface (VG), Xception (X), GoCaR (Go), OpenPose (OP), Facial Action Units (AU), and all visual feature set (aV) are used. Two types of alignment are used to a) align to eGeMAPS (GA) or b) aggregate on FastText word features (FA).

System	Features	Alig.	c-Valence F1 devel / test	c-Valence UAR devel / test	c-Valence Combined devel / test	c-Arousal F1 devel / test	c-Arousal UAR devel / test	c-Arousal Combined devel / test	Combined devel / test
Fine-tuned Albert	RT		36.18 / 34.21	33.17 / 33.05	35.16 / 33.81	33.33 / 37.14	33.69 / 34.30	33.45 / 36.18	34.30 / 35.00
LSTM + Self-ATT	DS	GA	34.17 / 34.60	34.07 / 35.00	34.13 / 34.74	38.03 / 37.54	38.43 / 36.78	38.17 / 37.28	36.15 / 36.01
LSTM + Self-ATT	eG	GA	33.26 / 34.44	32.16 / 33.94	32.89 / 34.27	34.39 / 33.33	34.44 / 32.87	34.41 / 33.18	33.65 / 33.73
LSTM + Self-ATT	FT	GA	38.41 / 36.19	37.75 / 36.22	38.18 / 36.20	35.15 / 34.92	35.78 / 37.10	35.37 / 35.66	36.78 / 35.93
LSTM + Self-ATT	eG + FT	GA	34.92 / 35.32	34.05 / 34.87	34.63 / 35.16	34.39 / 35.48	34.48 / 35.42	34.42 / 35.46	34.53 / 35.31
LSTM + Self-ATT	X	GA	36.21 / 36.83	35.75 / 36.61	36.06 / 36.75	40.38 / 35.16	40.51 / 34.87	40.43 / 35.06	38.24 / 35.91
LSTM + Self-ATT	aV	GA	35.61 / 34.92	35.10 / 34.41	35.44 / 34.75	38.11 / 34.21	38.26 / 35.39	38.16 / 34.61	36.80 / 34.68
LSTM + Self-ATT	eG + FT + aV	GA	36.06 / 37.14	35.20 / 37.14	35.77 / 37.14	39.92 / 35.16	40.44 / 34.76	40.10 / 35.02	37.93 / 36.08
MMT	FT + eG + X		38.28 / **39.92**	37.62 / **40.52**	38.06 / **40.12**	41.87 / 37.30	40.83 / 37.87	41.52 / 37.50	39.79 / **38.81**
MMT	FT + eG + VG		37.38 / 32.78	38.19 / 32.53	37.65 / 32.69	**47.12** / 41.19	**45.55** / 39.01	**46.58** / 40.45	42.12 / 36.57
MMT	DS + eG + VG		39.40 / 32.54	38.08 / 32.40	38.95 / 32.49	45.77 / 41.03	44.66 / **40.63**	45.39 / 40.89	**42.17** / 36.69
MMT	X + eG + VG		38.28 / 36.43	37.76 / 37.39	38.10 / 36.76	45.24 / 40.95	43.81 / 38.66	44.76 / 40.17	41.43 / 38.46
MMT	FT + eG + AU		36.93 / 39.92	37.35 / 39.57	37.07 / 39.80	43.15 / 34.76	41.88 / 34.87	42.72 / 34.80	39.89 / 37.30
MMT	FT + eG + OP		**39.48** / 38.81	**39.17** / 38.64	**39.37** / 38.75	38.88 / 37.70	38.95 / 38.10	38.90 / 37.83	39.14 / 38.29
MMT	OP + eG + AU		37.30 / 36.67	36.34 / 37.45	36.97 / 36.93	43.15 / 34.68	42.01 / 35.69	42.76 / 35.03	39.87 / 35.98
End2You	FT + eG + X		37.19 / 33.54	35.70 / 33.18	36.68 / 33.42	42.76 / 32.45	42.67 / 33.34	42.73 / 32.75	39.70 / 33.08
SVM	eG		36.33 / 33.10	34.79 / 34.13	35.81 / 33.45	43.52 / 34.37	42.27 / 33.43	43.10 / 34.05	39.45 / 33.75
SVM	DS		34.08 / 34.29	33.21 / 34.07	33.79 / 34.21	41.35 / 42.30	40.18 / 40.18	40.18 / 41.83	36.98 / 38.02
SVM	X		38.28 / 37.94	37.09 / 37.94	37.87 / 37.94	46.22 / 41.35	45.25 / 40.52	45.89 / 41.07	41.88 / 39.50
SVM	VG		37.08 / 32.94	37.01 / 32.63	37.06 / 32.83	46.44 / **42.46**	45.21 / **43.07**	46.02 / **42.67**	41.54 / 37.75
SVM	FT		37.90 / 36.43	36.00 / 35.37	37.26 / 36.07	45.17 / 38.25	44.53 / 39.67	44.95 / 38.74	41.10 / 37.40

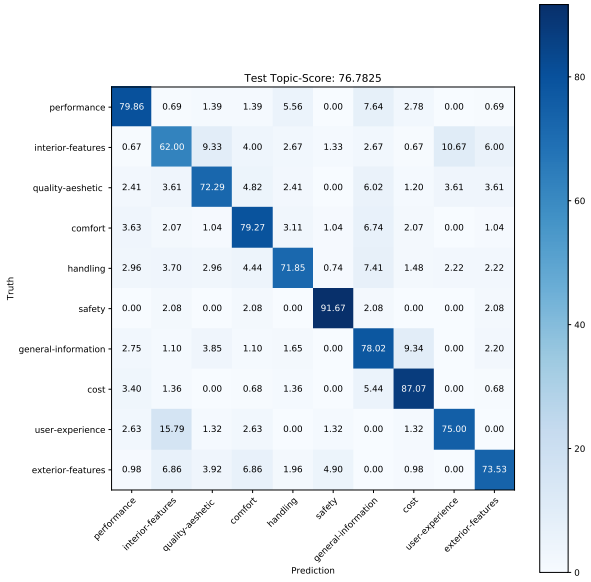

Figure 2: Relative confusion matrix over all 10 topics of fine-tuned Albert (left) as well as the MMT (FastText, eGeMAPS and Xception) for the prediction of valence (middle) and MMT (FastText, eGeMAPS and Xception) for the prediction of arousal (right) classes on the test partition for sub-challenges MuSe-Topic.

classes) as well as three classes (low, medium and high) of valence and arousal have to be predicted from video parts containing the discussed topic; and, iii) MuSe-Trust, where the level of continuous trustworthiness has to be predicted from features and/or affective annotations. By intention, we decided to use open-source software to extract a wide range of feature sets to deliver the highest possible transparency and realism for the baselines. Besides the features, we also share the raw data and the developed code for our baselines on a public platform. Results indicate that: i) the level of affection in-the-wild is best predicted when the system is trained on the the raw audio features; ii) for MuSe-Topic, (NLP-specific) Transformers are clearly superior when it comes to the prediction of topics, and no system is clearly outperforming on the three class valence and arousal prediction; and iii), in MuSe-Trust, adding valence and arousal contours as 'signals' in addition to other features is beneficial for the prediction of trustworthiness. The baselines also show the challenge ahead in mastering multimodal sentiment analysis, in particular when data are collected in user-generated, noisy environments. In the participants' and future efforts, we expect novel exciting combinations of the modalities – potentially also such as linking modalities on earlier stages or more closely.

8 ACKNOWLEDGMENTS

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), the EPSRC Grant No. 2021037, and the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B). We thank the sponsors of the Challenge BMW Group and audEERING.

REFERENCES

[1] Luis Aguado, Francisco J Román, María Fernández-Cahill, Teresa Diéguez-Risco, and Verónica Romero-Ferreiro. 2011. Learning about faces: effects of trustworthiness on affective evaluation. *The Spanish journal of psychology* 14, 2 (2011), 523–534.

[2] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn W Schuller. 2017. Snore Sound Classification Using Image-Based Deep Spectrum Features.. In *INTERSPEECH*, Vol. 434. 3512–3516.

[3] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2020. Gated multimodal networks. *Neural Computing and Applications* (2020), 1–20.

[4] T. Baltrušaitis, P. Robinson, and L. Morency. 2016. OpenFace: an Open Source Facial Behavior Analysis Toolkit. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, Lake Placid, NY. 10 pages.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2017. VGGFace2: A dataset for recognising faces across pose and age. *CoRR* abs/1710.08092 (2017). arXiv:1710.08092 http://arxiv.org/abs/1710.08092

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[8] Nicholas Cummins, Shahin Amiriparian, Gerhard Hagerer, Anton Batliner, Stefan Steidl, and Björn W Schuller. 2017. An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings of the 25th ACM international conference on Multimedia*. 478–484.

[9] Maria Teresa Cuomo, Debora Tortora, Alex Giordano, Giuseppe Festa, Gerardino Metallo, and Erika Martinelli. 2020. User-generated content in the era of digital well-being: A netnographic analysis in a healthcare marketing context. *Psychology & Marketing* (2020).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[12] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[13] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *The IEEE Winter Conference on Applications of Computer Vision*. 1470–1478.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Dimitrios Kollias, Attila Schulc, Elnar Hajiyev, and Stefanos Zafeiriou. 2020. Analysing affective behavior in the first ABAW 2020 competition. *arXiv preprint arXiv:2001.11409* (2020).

[16] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2019. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision* (2019), 1–23.

[17] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1eA7AEtvS

[18] I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* (1989), 255–268.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[22] Saif M Mohammad. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*. Elsevier, 201–237.

[23] Vedhas Pandit and Björn Schuller. 2019. On Many-to-Many Mapping Between Concordance Correlation Coefficient and Mean Square Error. *arXiv preprint arXiv:1902.05180* (2019).

[24] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, Mark W. Jones Xianghua Xie and Gary K. L. Tam (Eds.). BMVA Press, Article 41, 12 pages. https://doi.org/10.5244/C.29.41

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[26] Frits K Pil and Matthias Holweg. 2004. Linking product variety to order-fulfillment strategies. *Interfaces* 34, 5 (2004), 394–403.

[27] Daniel Preoţiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*. 9–15.

[28] Xiaoyu Qiu, Zhiquan Feng, Xiaohui Yang, and Jinglan Tian. 2020. Multimodal Fusion of Speech and Gesture Recognition based on Deep Learning. In *Journal of Physics: Conference Series*, Vol. 1453. 012092.

[29] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[30] 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.

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

[32] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[33] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

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

[35] 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).

[36] Björn W Schuller, Stefan Steidl, Anton Batliner, Peter B Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian B Pokorny, Eva-Maria Rathner, Katrin D Bartl-Pokorny, et al. 2018. The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats.. In *Interspeech*. 122–126.

[37] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[38] 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.

[39] Lukas Stappen, Xinchen Du, Vincent Karas, Stefan Müller, and Björn W Schuller. 2020. Go-CaRD–Generic, Optical Car Part Recognition and Detection: Collection, Insights, and Applications. *arXiv preprint arXiv:2006.08521* (2020).

[40] Lukas Stappen, Vincent Karas, Nicholas Cummins, Fabien Ringeval, Klaus Scherer, and Björn Schuller. 2019. From speech to facial activity: towards cross-modal sequence-to-sequence attention networks. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.

[41] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology* 61, 12 (2010), 2544–2558.

[42] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. *CoRR* abs/1906.00295 (2019). arXiv:1906.00295 http://arxiv.org/abs/1906.00295

[43] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing* 11, 8 (2017), 1301–1309.

[44] Panagiotis Tzirakis, Stefanos Zafeiriou, and Björn Schuller. 2019. Real-world automatic continuous affect recognition from audiovisual signals. In *Multimodal Behavior Analysis in the Wild*. Elsevier, 387–406.

[45] Panagiotis Tzirakis, Stefanos Zafeiriou, and Bjorn W Schuller. 2018. End2You–The Imperial Toolkit for Multimodal Profiling by End-to-End Learning. *arXiv preprint arXiv:1802.01115* (2018).

[46] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. 2018. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5089–5093.

[47] 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.

[48] Zheng Wang, Jie Zhou, Jing Ma, Jingjing Li, Jiangbo Ai, and Yang Yang. 2020. Discovering attractive segments in the user-generated video streams. *Information Processing & Management* 57, 1 (2020), 102130.

[49] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2015. WIDER FACE: A Face Detection Benchmark. *CoRR* abs/1511.06523 (2015). arXiv:1511.06523 http://arxiv.org/abs/1511.06523

[50] Amir Zadeh, Paul Pu Liang, Louis-Philippe Morency, Soujanya Poria, Erik Cambria, and Stefan Scherer. 2018. Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML). In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*.

[51] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23 (04 2016). https://doi.org/10.1109/LSP.2016.2603342