Sentiment analysis and topic recognition in video transcriptions

Lukas Stappen, Alice Baird, Erik Cambria, Björn Schuller, Erik Cambria

Angaben zur Veröffentlichung / Publication details:

Stappen, Lukas, Alice Baird, Erik Cambria, Björn Schuller, and Erik Cambria. 2021. "Sentiment analysis and topic recognition in video transcriptions." *IEEE Intelligent Systems* 36 (2): 88–95. https://doi.org/10.1109/mis.2021.3062200.

Sentiment Analysis and Topic Recognition in Video Transcriptions

Lukas Stappen and Alice Baird, *University of Augsburg, Augsburg, 86159, Germany*

Erik Cambria, *Nanyang Technological University, 639798, Singapore*

Björn W. Schuller, *Imperial College London, London, SW7 2BU, U.K.*

Nowadays, videos are an integral modality for information sharing on the World Wide Web. However, systems able to automatically understand the content and sentiment of a video are still in their infancy. Linguistic information transported in spoken parts of a video is known to convey valuable properties in regards to context and emotions. In this article, we explore a lexical knowledge-based extraction approach to obtain such understanding from the video transcriptions of a large-scale multimodal dataset (MuSe-CAR). To this end, we use SenticNet to extract natural language concepts and fine-tune several feature types on a subset of MuSe-CAR. With these features, we explore the content of a video as well as learning to predict emotional valence, arousal, and speaker topic classes. Our best model improves the linguistic baseline from the MuSe-Topic 2020 subchallenge by almost 3% (absolute) for the prediction of valence on the predefined challenge metric and outperforms a variety of baseline systems that require much higher computational power than the one proposed herein.

The importance of video in social media and pure video-based platforms is rapidly growing. In 2018, the numbers of users on video-based platforms such as YouTube (+27%) and TikTok (+109%) has grown faster than the biggest social media platform Facebook (+11%).[1] Due to the increasing availability of computing power and machine learning techniques, new ways to explore video content are sought. Multimodal sentiment analysis (MSA) in-the-wild is one research area in which structure of these massive amounts of data can begin to be obtained.[2–4]

At its core, the MSA research field aims to understand (within a video) the schematic structure of sentiments including the sentiment holder, the emotional disposition, and the reference object.[5] The emotion is directed toward the reference objects. They occur in various granularity, for instance, a video might cover multiple topics and aspects. A topic can be interpreted as the announcement of a discourse, thus, the utterance of semantics beyond individual scenes and sentences. A video signal yields three modalities: the visual (e.g., facial expressions), the audio (e.g., vocal characteristics), and the textual information (transcription of the spoken word). It has been found, however, that the textual modality has the greatest impact in understanding the context (e.g., topic).[6]

We present a sentiment analysis study focusing on topics and emotions in video car reviews from YouTube within our contribution. Thereby, we aim at a more in-depth exploration of the spoken word, hence, the use of transcriptions. There are two common ways to utilize text computationally: a) understand the meaning of words from their symbolic representation through knowledge-based and statistical approaches; b) learning a continuous vector space (embeddings) from the symbolic space of words. Where the first often focuses on the construction of a taxonomy and the second is based on neural learning. Instead, we extract high-level natural language concepts as features and apply them in a specific domain without constructing a new taxonomy. The, therefore, necessary vectorial representations can be obtained by subsymbolic AI frameworks based on commonsense computing.[7]

A very popular theoretic grounding is the Hourglass of Emotions,[8] a biologically inspired and psychologically motivated emotion categorization model for sentiment analysis. Building on the categorization provided by this model, SenticNet[9] is a commonsense knowledge based that provides a set of semantics, sentics, and polarity associated with natural language concepts. We utilize SenticNet to extract the aforementioned attributes and transform them into domain-specific features. Overall, within the study, we provide the following two key contributions.

- Gain a better understanding of the usefulness of high-contextual features to analyze transcriptions—an audio codepending modality—as groundwork for the addition to raw signals (visual, audio, and text) normally used for MSA since others do not contain any bottom–up and top–top understanding of our world. One exploratory way of doing so is the comparison of these to human annotations.
- We experiment with topic and emotion recognition from extracted features based on a subsymbolic framework. To do this, we use the transcriptions of MuSe-CaR, the largest English speaking MSA dataset.

RELATED WORK

Video is a versatile source of information for sentiment prediction. Utilizing the transcription of video utterances in combination with other modalities[10–12] classified the general sentiment. However, these approaches neglect that human communication is symbolic, naturally ordered in a hierarchical structure. In this respect, knowledge-based frameworks use relational multiword expressions to analyze text from other sources (excluding video transcriptions). Computational, dictionary-based analysis of content was first proposed by Stone *et al.*[13] WordNet-Affect contains almost 3,000 synsets, e.g., labels that indicate emotion and mood categories. Based on the Russell circumplex model of emotions, affective norms from English words examine the dimensions of valence, arousal, and dominance.[14] To date, SenticNet is the largest of these frameworks, containing 200,000 concepts, which maps a word to sentic and moodtag dimensions of the Hourglass of Emotions.

To truly understand the meaning of a sentiment directed at an aspect or topic, we need to contextualize it within the overarching underlying elements of our world, such as social norms. Manual aspect specification often comes along with intensive domain knowledge and expensive/time-consuming extraction. Therefore, automated sentiment and aspect extraction has been thoroughly studied using supervised and unsupervised algorithms for the past two decades.[15–19]

First efforts toward extracting topics and aspects came up in the supervised context of analyzing customer reviews and social media data using rule mining[20] and lexicons.[21] Aspect and topic extraction on closed captions from YouTube videos, containing customer reviews, has been previously implemented using an attention-based network.[22] However, the authors' approach only used a small dataset of seven videos and did not consider high-level natural language concept features in predicting sentiment or for aspect extraction. Besides this, caption mining was previously used in the context of scene segmentation,[23] video activity recognition,[24] and movie genre classification.[25] Targeting the automotive domain[26] demonstrated that emotional and factual knowledge, such as a sentiment of a product feature, can be extracted from text using a combination of lexical methods including SenticNet.

Concerning our current contribution, it is evident that work applying commonsense knowledge based or focusing on topics, compromising several sentences, in video transcriptions are clearly in the minority, and there have been minimal efforts in applying these features explicitly.

METHODOLOGY

A pipeline of the entire process is depicted in Figure 1. Beginning with the automatic transcription from spoken language, natural world concepts are obtained using SenticNet. The concepts are summarized to sentence- or segment-level features and, then, utilized for descriptive analysis or by machine learning methods to train models predicting the targets (topic, arousal, and valence). As mentioned previously with SenticNet, we can extract various features, including, sentics, moodtags, and semantics.

In this contribution, we utilize versions 5 and 6 of SenticNet. From version 5, we extract four sentics: pleasantness, attention, sensitivity, and aptitude. From version 6, the sentics are different, and so we extract: introversion, temper, attitude, and sensitivity. Primary and secondary moodtags can also be extracted. When observing the Hourglass model, the

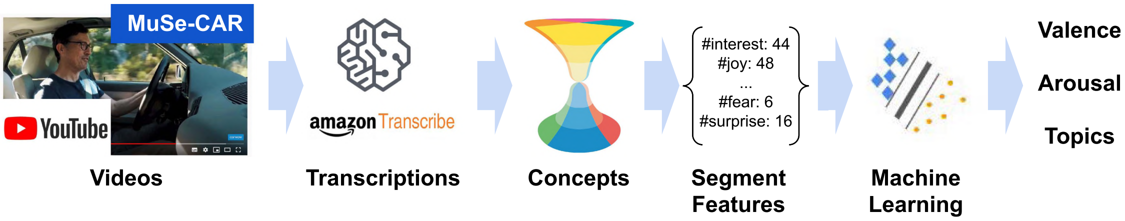


FIGURE 1. Overview of our processing pipeline from the raw video to predicting arousal, valence, and speech topics.

moodtags are below the sentics and include labels such as bliss, ecstasy, and delight. Semantics are also extracted, and these can be described as concept clusters that are semantically related to the segment and share a similar lexical function (cf. Figure 5 for examples).

Extraction of SenticNet Features: Sentic API[†] serves as our core feature extractor to obtain the semantics, sentics, moodtags, and polarity for each n-gram of our corpus. We apply a very simplistic data cleaning, removing stopwords, such as personal pronouns (e.g., I, me, you, him), articles (the, a, an), and conjunctions (e.g., and, or). To utilize these in sentence or segment context, they have to be aggregated. Formally, given a sequence n of words $\bar{w}_s = [w_1, \ldots, w_n]$ for a segment s, we receive a sequence m of concepts, which can vary in length to $\bar{w}_s$, $\bar{c}_s = [c_1, \ldots, c_m]$, which are embedded to a vector h_s. For the discrete concepts of semantics and sentics, h_s is a concatenation that forms an n-hot encoded vector. For the concepts that come with a continuous-valued intensity, such as moodtags and polarity, the changing length n of the respective context s has to be considered. We do this by applying a normalized average over the measurements across the context.

Modeling: In machine learning, a straightforward method to evaluate features' predictive power is through the usage of support vector machines (SVMs). The robustness against high-dimensional feature data is also an advantageous property when dealing with n-hot encoded vectors. In the past, SVMs showed results close to or better than other state-of-the-art algorithms in similar settings, such as neural networks,[27] particularly when there is not a massive amount of data available.[28] For our experiments, we employ a linear SVM classifier implementation from the python package SCIKIT-LEARN. We predict our targets y_s from our concept vector $\bar{c}_s$ or feature h_s without applying normalization. The C value is tuned from 10^{-5} to 1 on the development set using 10,000 iterations and the best used for the prediction on the test set.

In contrast to our SVM approach, we also intend to fine-tune the extracted semantic concepts $\bar{c}_s$ to domain-specific embeddings h_s using neural learning. Similar to the word embeddings training concept, we assign every semantic a fixed position in a one-hot encoded vector. These sparse input vectors are then compressed to a 100-dimensional embedding space: $h_s = \sigma(\bar{c}_s)$, where the σ layer has a sigmoid activation function. Based on the embedding vectors, additional layers can build upon this, which condense the information into a single meaningful vector, representing an entire sentence or segment. The summary vector either predicts the target directly or is used as a feature vector for an SVM. To improve generalization and promote independence between feature maps, we utilize embedding dropout to drop single features in the embedding space and time-step dropout to drop entire embeddings instead of individual features.

DATASET: THE MUSE-TOPIC SUBCHALLENGE

In this section, we describe the dataset and the prediction task environment. The MuSe-CaR dataset[29] is a large, multimodal dataset focused on sentiment modeling in automotive video reviews supporting various research directions with predefined data subsets holding unique, task-specific properties and labels. In this article, we utilize the MuSe-Topic subset.[‡] This subchallenge was released as part of the MuSe 2020 challenge[27,30] and provides 10-classes of domain-specific speaker topics as the target of three classes (low, medium, and high) of valence and arousal emotions.

Although the subset provides multiple modalities, in this article, we only focus on the language modality, disregarding audio–visual signals. In recent years, speech-to-text services improved drastically, reaching almost human-level quality in the English language. In order to receive transcriptions even capturing

[†] https://sentic.net/api

[‡] Download available at https://zenodo.org/record/4134733

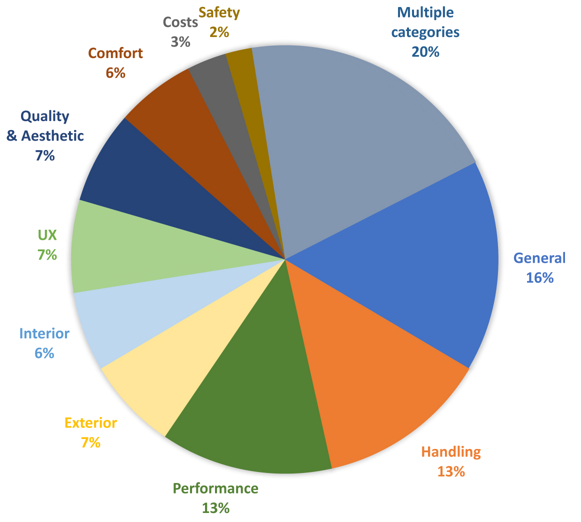


FIGURE 2. Relative distribution of the segments regarding speaker topics.

domain-specific vocabulary (e.g., "eDrive"), Stappen *et al.*[29] created a customized dictionary of typical automotive terms. The Amazon Transcribe[§] service enables the use of such dictionaries. The transcribed audio signals include full punctuation, resulting in a total of 28,295 sentences.

The content can be divided into several segments. A segment always comprises of only one topic. A topic is defined as the vocalization made by the reviewer about a group of homogeneous conversation subjects. For example, interior features include diverse information of entities, functions, and aspects inside a vehicle, such as the infotainment system and device connectivity. One speaker topic segment often consists of one or multiple sentences. However, as in the subchallenge, we excluded around 20% of sentences that belong to multiple segment topic labels. An overview of the topics and distribution is depicted in Figure 2. In addition, for each topic segment, one valence and one arousal class are given. The classes represent the mean value of the temporally aggregated continuous annotations divided into three equally sized classes (33%) for each label.

As in this subchallenge, we report the weighted score combining unweighted average recall (UAR) and F1 (micro) measures independently for each prediction (valence, arousal, and topic). In addition, we use the same training, development, and test partitions to enable a fair comparison to the text-based baseline models from the work by Stappen *et al.*[27]

EXPLORATORY ANALYSIS

Speaker topics: First, we make use of the concepts of semantics to obtain a contextual description of the videos. Figure 3 exemplary compares the semantics for some topic extracted by the two versions. We exclude concepts that frequently reappear across all classes (one-against-all) to focus on the most distinctive concepts for each class, which do not occur in most others. Although the found semantics differ between versions, they describe seemingly natural aspects of the topics. For our entire corpus, SenticNet version 5 provides 14,685 and version 6 provides 8,577 semantics. These extractions serve as a first indicator of the characteristic properties of the video content.

Emotions: Second, we explore the emotional information as a target of the topics, which also can be obtained by SenticNet in an unsupervised fashion. Figure 4 compares the distribution of the real, hence manually labeled by human raters, continuous arousal, and valence annotations, before the label aggregation to classes, to the SenticNet polarity extraction. While the Gaussian-shaped distribution of arousal values is almost entirely centering around 0, valence and the SenticNet output are skewed toward the spectrum's positive end. The version 6 extractions appear to be even more similar to the original valence annotations, due to broader and flatter distribution. This observed similarity of valence and sentiment polarity is well in line with previous research.[27] Overall, when it comes to interpreting videos regarding valence, SenticNet poses a strong indicator to harvest sentiment information from video transcripts without any additional annotations.

PREDICTION RESULTS

After applying SenticNet for the video transcriptions' explorative analysis, we also want to use the extracted features for the prediction tasks.

Speaker topics: Table 1 shows the results of the task of topic prediction. The naive *n*-hot encoded features combined with an SVM yields the best performance of all evaluated systems achieving a combined challenge metric of 56.18% on the development and 66.16% on the test set. As illustrated in Figure 5, the prediction errors are equally spread, with slightly more confusion between the interior and aesthetics classes as well as the cost and general information. Our domain-specific encoding falls slightly behind with 56.71% on the test set. We also evaluated a pure neural network architecture, further temporally encoding the embeddings using an LSTM, however, achieved slightly worse results.

[§] https://aws.amazon.com/transcribe/

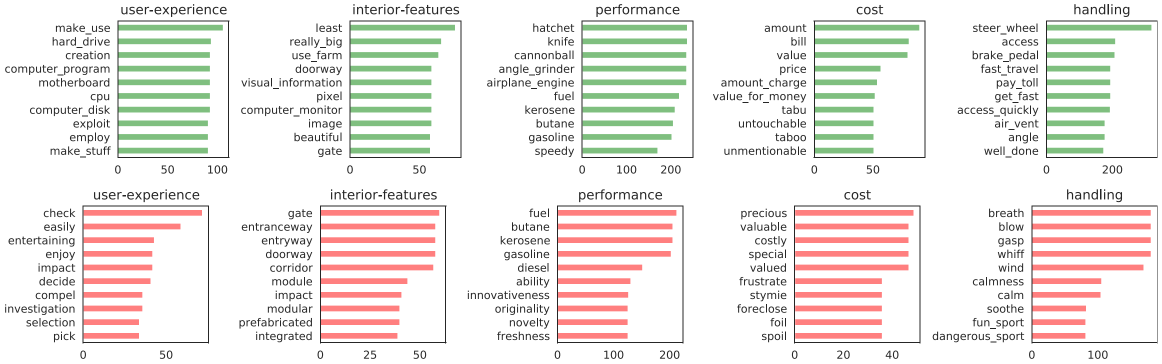


FIGURE 3. Number of occurrences (on sentence-level) of the top 10 semantics of 5 exemplary topics extracted from SenticNet 5 (green) and 6 (red), excluding the most common 100 semantics occurring also in other topics.

Compared to the baseline, the results appear very competitive. They outperform the LSTM with self-attention by more than 30%, even outperforming the multimodal transformer from the work by Stappen *et al.*[27] by almost 15%. Only Albert, to date, the most robust end-2-end NLP transformer for supervised NLP tasks exceeds our performance. This result was to be expected, given the considerable number of model parameters and the extensive pretraining on downstream NLP tasks with masses of text data available.

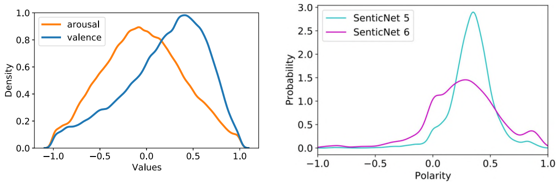


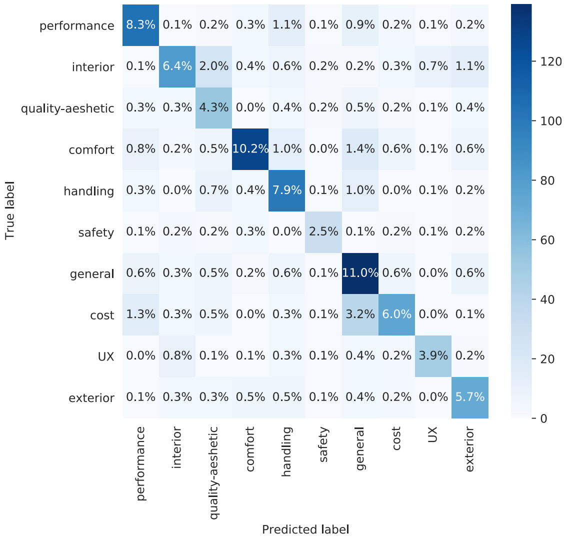
FIGURE 4. Density estimation of the continuously annotated dimensions of arousal (orange) and valence (blue) on the left and the SenticNet polarity intensities of SenticNet 5 (cyan) and SenticNet 6 (magenta) on the right.

TABLE 1. Muse-Topic: Reporting the combined score $(0.66 \cdot F1 + 0.34 \cdot UAR)$ for the prediction of topics using semantics on the devel(opment) and test set.

System		devel	test
Ours based on SenticNet 5	SVM: *n*-hot encoded	56.18	66.15
	SVM: NN embeddings	47.08	56.71
Ours based on SenticNet 6	SVM: *n*-hot encoded	46.22	57.09
	SVM: NN embeddings	40.67	49.01
baselines[27]	LSTM + Self-ATT: FT	21.44	36.20
	MMT: FT + eG + AU	44.33	52.98
	Albert: text	70.62	**76.78**

The baselines use FastText (FT), eGemaps (eG), and Facial Action Units (AU) as feature sets. Bold data highlights best performing system.

Emotions: Generally, Table 2 shows a clear advantage of SenticNet 6 over 5 for the predicting the emotion classes. The best results improve the baseline by almost 3% (combined score) absolute using sentics, moodtags, and polarity for valence. While the moodtags features seem slightly superior to the conceptual sentics, all are profiting from fusion. However, the picture is different when predicting arousal, where all



	performance	interior	quality-aeshetic	comfort	handling	safety	general	cost	UX	exterior
performance	8.3%	0.1%	0.2%	0.3%	1.1%	0.1%	0.9%	0.2%	0.1%	0.2%
interior	0.1%	6.4%	2.0%	0.4%	0.6%	0.2%	0.2%	0.3%	0.7%	1.1%
quality-aeshetic	0.3%	0.3%	4.3%	0.0%	0.4%	0.2%	0.5%	0.2%	0.1%	0.4%
comfort	0.8%	0.2%	0.5%	10.2%	1.0%	0.0%	1.4%	0.6%	0.1%	0.6%
handling	0.3%	0.0%	0.7%	0.4%	7.9%	0.1%	1.0%	0.0%	0.1%	0.2%
safety	0.1%	0.2%	0.2%	0.3%	0.0%	2.5%	0.1%	0.2%	0.1%	0.2%
general	0.6%	0.3%	0.5%	0.2%	0.6%	0.1%	11.0%	0.6%	0.0%	0.6%
cost	1.3%	0.3%	0.5%	0.0%	0.3%	0.1%	3.2%	6.0%	0.0%	0.1%
UX	0.0%	0.8%	0.1%	0.1%	0.3%	0.1%	0.4%	0.2%	3.9%	0.2%
exterior	0.1%	0.3%	0.3%	0.5%	0.5%	0.1%	0.4%	0.2%	0.0%	5.7%

FIGURE 5. Relative confusion matrix over all 10 speaker topics using our trained model based on SenticNet 5 *n*-hot encoded embeddings on the test partition.

TABLE 2. Muse-Topic: Reporting UAR, F1, and combined ($0.66 \cdot F1 + 0.34 \cdot$ UAR) for the prediction of valence and arousal using sentics, moodtags (mood), and polarity (pol) as well as fasttext as feature sets, which are feed into an SVM.

System	Feature(s)	Valence			Arousal		
		F1	UAR	Combined	F1	UAR	Combined
Ours based on SenticNet 5	sentics	35.66 / 35.16	34.65 / 35.17	35.32 / 35.16	33.56 / 33.73	34.51 / 33.69	33.88 / 33.72
	mood	33.78 / 37.86	33.53 / 38.04	33.70 / 37.92	36.78 / 35.79	36.30 / 33.50	36.62 / 35.01
	sentics + mood	34.38 / 37.78	34.22 / 37.89	34.33 / 37.82	35.96 / 35.63	35.49 / 33.60	35.80 / 34.94
	sentics + mood + pol	34.53 / 38.41	34.40 / 38.55	34.49 / 38.46	35.51 / 36.51	35.15 / 34.54	35.39 / 35.84
Ours based on SenticNet 6	sentics	35.51 / 36.59	34.77 / 36.72	35.26 / 36.63	32.43 / 35.16	33.45 / 35.91	32.78 / 35.42
	mood	38.65 / 36.83	38.28 / 37.57	38.52 / 37.08	35.66 / 38.02	35.33 / 35.86	35.55 / 37.29
	sentics + mood	38.13 / 38.57	37.89 / **38.90**	38.05 / 38.68	36.18 / 38.02	35.68 / 35.90	36.01 / 37.30
	sentics + mood + pol	37.68 / **38.65**	37.54 / 38.88	37.63 / **38.73**	35.96 / **38.33**	35.50 / 36.12	35.80 / 37.58
baseline[27]	FastText	37.90 / 36.43	36.00 / 35.37	37.26 / 36.07	45.17 / 38.25	44.53 / **39.67**	44.95 / **38.74**

Bold data highlights best performing system.

configuration performs worse than the text-embedding, low-level baseline.

These results are conclusive considering the exploratory observations of the emotions. In combination, they lead us to assume that the high-level contextual SenticNet features might be valuable, mainly, as emotional valence can be a challenge for unimodal audio-based approaches. It could also be fused with low-level text embeddings. Nevertheless, since the categories were derived from continuous signals and not directly labeled, further research should be conducted.

CONCLUSIONS

In this article, we explored subsymbolic representations gained from sentic concepts to gain insights into the emotional and contextual information provided by video transcriptions. Furthermore, we have successfully leveraged the derived features to automatically classify video segments regarding arousal and valence as well as 10 domain-specific speaker topics. In the future, one should build upon these promising results, using the semantics in a more unsupervised way to explore the content of videos by clustering and in combination with high-level feature sets of other modalities (e.g., face and voice features) for multimodal modeling.

REFERENCES

1. E. Ortiz-Ospina, "The rise of social media," Sep. 2019. [Online]. Available: https://ourworldindata.org/rise-of-social-media
2. M. Soleymani *et al.*, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, 2017, pp. 3–14.
3. A. Zadeh *et al.*, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
4. E. Cambria *et al.*, "Sentic blending: Scalable multimodal fusion for continuous interpretation of semantics and sentics," in *Proc. IEEE Symp. Comput. Intell. Human-Like Intell.*, 2013, pp. 108–117.
5. M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 101–111, Apr.–Jun. 2014.
6. I. Chaturvedi *et al.*, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognit. Lett.*, vol. 125, no. 264–270, 2019.
7. E. Cambria *et al.*, "Common sense computing: From the society of mind to digital intuition and beyond," in *Biometric ID Management and Multimodal Communication* (*Lecture Notes in Comput. Sci.*, 5707), J. Fierrez , Eds. Berlin, Germany: Springer, 2009, pp. 252–259.

8. Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria, "The Hourglass model revisited," *IEEE Intell. Syst.*, vol. 35, no. 5, pp. 96–102, Sep./Oct. 2020.
9. E. Cambria *et al.*, "SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, 2020, pp. 105–114.
10. L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 169–176.
11. V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 973–982.
12. N. Majumder *et al.*, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, 2018.
13. P. J. Stone, D. C. Dunphy, and M. S. Smith, *The General Inquirer: A Computer Approach to Content Analysis.* Cambridge, MA, USA: MIT Press, 1966.
14. M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," *Tech. Rep., Center Res. Psychophysiology*, Gainesville, Fl, USA, 1999.
15. N. Jakob and I. Gurevych, "Extracting opinion targets in a single-and cross-domain setting with conditional random fields," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2010, pp. 1035–1045.
16. Z. Chen, A. Mukherjee, and B. Liu, "Aspect extraction with automated prior knowledge learning," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 347–358.
17. T. A. Rana and Y.-N. Cheah, "Aspect extraction in sentiment analysis: Comparative analysis and survey," *Artif. Intell. Rev.*, vol. 46, no. 4, pp. 459–483, 2016.
18. S. Poria *et al.*, "Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis," in *Proc. Int. Joint Conf. Neural Netw.*, 2016, pp. 4465–4473.
19. Y. Ma *et al.*, "Sentic LSTM: A hybrid network for targeted aspect-based sentiment analysis," *Cogn. Comput.*, vol. 10, no. 4, pp. 639–650, 2018.
20. M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 168–177.
21. Q. Su *et al.*, "Using pointwise mutual information to identify implicit features in customer reviews," *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, Y. Matsumoto Eds. Berlin, Germany: Springer, 2006, pp. 22–30.
22. E. Marrese-Taylor, J. Balazs, and Y. Matsuo, "Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN," in *Proc. 8th Workshop Comput. Approaches Subjectivity, Sentiment, Social Media Anal.*, 2017, pp. 102–111.
23. S. Gupta and R. J. Mooney, "Using closed captions to train activity recognizers that improve video retrieval," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2009, pp. 30–37.
24. S. Gupta and R. Mooney, "Using closed captions as supervision for video activity recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2010, pp. 1083–1088.
25. D. Brezeale and D. J. Cook, "Using closed captions and visual features to classify movies by genre," in *Proc. 7th Int. Workshop Multimedia Data Mining*, 2006.
26. A. Weichselbraun, S. Gindl, F. Fischer, S. Vakulenko, and A. Scharl, "Aspect-based extraction and analysis of affective knowledge from social media streams," *IEEE Intell. Syst.*, vol. 32, no. 3, pp. 80–88, May/Jun. 2017.
27. L. Stappen *et al.*, "MuSe 2020 Challenge and Workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in *Proc. 1st Int. Multimodal Sentiment Anal. Real-Life Media Challenge Workshop*, 2020, pp. 35–44.
28. B. W. Schuller *et al.*, "The INTERSPEECH 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proc. Interspeech*, 2020, pp. 2042–2046.
29. L. Stappen *et al.*, "The multimodal sentiment analysis in car reviews (MuSe-CaR) Dataset: Collection, insights and improvements," 2021.
30. L. Stappen *et al.*, "Summary of MuSe 2020: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 4769–4770.

LUKAS STAPPEN is currently working toward the Ph.D. degree with the Chair for Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany. He is also a Ph.D. Fellow of the BMW Group, Munich, Germany. His research interests include affective computing, multimodal sentiment analysis, and multimodal representation learning with a core focus on "in-the-wild" environments. He received the M.Sc. degree in data science with distinction from King's College London, London, U.K., in 2017. He then joined the group for machine learning in health informatics. He is the corresponding author of this article. Contact him at stappen@ieee.org.

ALICE BAIRD is currently a Ph.D. Fellow of the Center for Digitization Bavaria (ZD.B), supervised by Prof. Björn Schuller with the Chair of Embedded Intelligence for Healthcare and Wellbeing, University of Augsburg, Augsburg, Germany. Her research is focused on intelligent audio analysis in the domain of speech and general audio, and her interests include health informatics, affective computing, and computational paralinguistics. She received the M.F.A. degree in sound art from the Computer Music Center, Columbia University, New York, NY, USA. Contact her at alice.baird@informatik.uni-augsburg.de.

ERIK CAMBRIA is currently an Associate Professor with Nanyang Technological University, Singapore. His main research interests include AI and affective computing. He received the Ph.D. degree in computing science and mathematics through a joint program between the University of Stirling, Stirling, U.K., and MIT Media Lab, Cambridge, MA, USA. Contact him at cambria@ntu.edu.sg.

BJÖRN W. SCHULLER is currently an Adjunct Teaching Professor in machine intelligence and signal processing with TUM, Munich, Germany. He is also a Full Professor of artificial intelligence and the Head of GLAM with Imperial College London, London, U.K., a Full Professor of embedded intelligence for health care and wellbeing with the University of Augsburg, Augsburg, Germany, among other professorships and affiliations. He received the diploma, doctoral degree, and habilitation from TUM. He has (co)authored more than 1,000 publications. He is a Fellow of the IEEE and Golden Core Awardee of the IEEE Computer Society, Fellow of the BCS, Fellow of the ISCA, President-Emeritus of the Editor-in-Chief of the *IEEE Transactions on Affective Computing* among other commitments. Contact him at bjoern.schuller@imperial.ac.uk.