Abstract—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 work, we explore a lexical knowledge-base 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 sub-challenge by almost 3 % (absolute) for the prediction of valence on the predefined challenge metric and outperforms a variety of baseline systems which require much higher computational power than the one proposed herein.
Affective Computing and Sentiment Analysis

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 towards 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-base 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 sub-symbolic 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 above-mentioned 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-down 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 sub-symbolic 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], [11], [12] classified the general sentiment. However, these approaches neglect that human communication is symbolic, naturally ordered in a hierarchical structure. In this respect, knowledge-base frameworks use relational multiword expressions to analyze text from other sources (excluding video transcriptions). Computational, dictionary-based analysis of content was first proposed by [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 examines 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], [16], [17], [18], [19].
First efforts towards 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 7 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 base 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, valence). As mentioned previously with SenticNet, we can extract various features, including, sentsics, moodtags and semantics.

In this contribution, we utilize versions 5 and 6 of SenticNet. From version 5, we extract four sentsics: pleasantness, attention, sensitivity, and aptitude. From version 6, the sentsics 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 moodtags, are below the sentsics 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\(^1\) serves as our core feature extractor, to obtain the semantics, sentsics, 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 \( \bar{w}_s = [w_1, ..., w_n] \) for a segment \( s \), we receive a sequence \( m \) of concepts, which can vary in length to \( \bar{c}_s = [c_1, ..., c_m] \), which are embedded to a vector \( h_s \). For the discrete concepts of semantics and sentsics, \( h_s \) is a concatenation which 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

---

\(^1\)https://sentic.net/api
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. Similarly 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 $\sigma$ 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 SUB-CHALLENGE**

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 work, 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 work 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 domain-specific vocabulary (e.g., “eDrive”), the authors of [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 which belong to multiple segment topic labels. An overview of the topics and distribution are 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.

---

2 Download available at https://zenodo.org/record/4134733
3 https://aws.amazon.com/transcribe/
As in this sub-challenge, 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 [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 which 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 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 labelled 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 2 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
Table 1. MuSe-Topic: Reporting UAR, F1, and Combined (0.66 · F1 + 0.34 · 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.

<table>
<thead>
<tr>
<th>System</th>
<th>Feature(s)</th>
<th>Valence UAR</th>
<th>F1</th>
<th>Combined</th>
<th>Arousal UAR</th>
<th>F1</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours based on SenticNet 5</td>
<td>sentics</td>
<td>35.66 / 35.16</td>
<td>34.66 / 35.17</td>
<td>35.32 / 35.16</td>
<td>33.56 / 33.73</td>
<td>34.17 / 33.69</td>
<td>33.88 / 33.72</td>
</tr>
<tr>
<td></td>
<td>mood</td>
<td>33.78 / 37.86</td>
<td>33.53 / 38.04</td>
<td>33.70 / 37.92</td>
<td>36.78 / 35.79</td>
<td>36.30 / 33.50</td>
<td>36.62 / 35.01</td>
</tr>
<tr>
<td></td>
<td>sentics + mood</td>
<td>34.38 / 37.78</td>
<td>34.22 / 37.89</td>
<td>34.33 / 37.82</td>
<td>35.96 / 35.63</td>
<td>35.49 / 33.60</td>
<td>35.80 / 34.94</td>
</tr>
<tr>
<td></td>
<td>sentics + mood + pol</td>
<td>34.53 / 38.41</td>
<td>34.40 / 38.55</td>
<td>34.49 / 38.46</td>
<td>35.51 / 36.51</td>
<td>35.15 / 34.54</td>
<td>35.39 / 35.84</td>
</tr>
<tr>
<td>Ours based on SenticNet 6</td>
<td>sentics</td>
<td>35.51 / 36.59</td>
<td>34.77 / 36.72</td>
<td>35.26 / 36.63</td>
<td>32.43 / 35.16</td>
<td>33.43 / 35.91</td>
<td>32.78 / 35.42</td>
</tr>
<tr>
<td></td>
<td>mood</td>
<td>38.65 / 38.83</td>
<td>38.28 / 37.57</td>
<td>38.52 / 37.08</td>
<td>35.66 / 38.02</td>
<td>35.33 / 35.86</td>
<td>35.55 / 37.29</td>
</tr>
<tr>
<td></td>
<td>sentics + mood</td>
<td>38.13 / 38.57</td>
<td>37.89 / 38.90</td>
<td>38.05 / 38.68</td>
<td>36.18 / 38.02</td>
<td>35.68 / 35.90</td>
<td>36.01 / 37.30</td>
</tr>
<tr>
<td></td>
<td>sentics + mood + pol</td>
<td>37.68 / 38.65</td>
<td>37.54 / 38.88</td>
<td>37.63 / 38.73</td>
<td>35.96 / 38.33</td>
<td>35.50 / 36.12</td>
<td>35.80 / 37.58</td>
</tr>
<tr>
<td>baseline [27]</td>
<td>FastText</td>
<td>37.90 / 36.43</td>
<td>36.00 / 35.37</td>
<td>37.26 / 36.07</td>
<td>45.17 / 38.25</td>
<td>44.55 / 39.67</td>
<td>44.95 / 38.74</td>
</tr>
</tbody>
</table>

Table 2. MuSe-Topic: Reporting the combined score (0.66 · F1 + 0.34 · UAR) for the prediction of topics using semantics on the develop(opment) and test set. The baselines use FastText (FT), eGemaps (eG), and Facial Action Units (AU) as feature sets.

<table>
<thead>
<tr>
<th>System</th>
<th>Feature(s)</th>
<th>valence</th>
<th>test</th>
<th>valence</th>
<th>test</th>
<th>valence</th>
<th>test</th>
<th>valence</th>
<th>test</th>
<th>valence</th>
<th>test</th>
<th>valence</th>
<th>test</th>
<th>valence</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours based on SenticNet 5</td>
<td>SVM: n-hot encoded</td>
<td>56.18</td>
<td>66.15</td>
<td>39.46</td>
<td>40.27</td>
<td>38.90</td>
<td>39.67</td>
<td>38.50</td>
<td>39.16</td>
<td>36.05</td>
<td>38.33</td>
<td>38.05</td>
<td>38.68</td>
<td>36.56</td>
<td>38.02</td>
</tr>
<tr>
<td></td>
<td>SVM: NN embeddings</td>
<td>47.08</td>
<td>56.71</td>
<td>37.54</td>
<td>38.26</td>
<td>37.82</td>
<td>38.05</td>
<td>36.58</td>
<td>37.82</td>
<td>35.55</td>
<td>37.29</td>
<td>35.55</td>
<td>37.29</td>
<td>35.55</td>
<td>37.29</td>
</tr>
<tr>
<td>Ours based on SenticNet 6</td>
<td>SVM: n-hot encoded</td>
<td>46.22</td>
<td>57.09</td>
<td>40.07</td>
<td>40.76</td>
<td>39.71</td>
<td>40.07</td>
<td>39.53</td>
<td>40.07</td>
<td>39.07</td>
<td>40.07</td>
<td>39.07</td>
<td>40.07</td>
<td>39.07</td>
<td>40.07</td>
</tr>
<tr>
<td></td>
<td>SVM: NN embeddings</td>
<td>40.67</td>
<td>49.01</td>
<td>39.07</td>
<td>39.71</td>
<td>39.53</td>
<td>40.07</td>
<td>39.07</td>
<td>40.07</td>
<td>39.07</td>
<td>40.07</td>
<td>39.07</td>
<td>40.07</td>
<td>39.07</td>
<td>40.07</td>
</tr>
<tr>
<td>baselines [27]</td>
<td>MMT: FT + eG + AU</td>
<td>21.44</td>
<td>36.20</td>
<td>24.33</td>
<td>32.98</td>
<td>23.68</td>
<td>32.33</td>
<td>23.00</td>
<td>31.66</td>
<td>22.33</td>
<td>31.00</td>
<td>21.33</td>
<td>30.00</td>
<td>21.33</td>
<td>30.00</td>
</tr>
<tr>
<td></td>
<td>Albert: text</td>
<td>70.62</td>
<td>76.78</td>
<td>70.62</td>
<td>76.78</td>
<td>70.62</td>
<td>76.78</td>
<td>70.62</td>
<td>76.78</td>
<td>70.62</td>
<td>76.78</td>
<td>70.62</td>
<td>76.78</td>
<td>70.62</td>
<td>76.78</td>
</tr>
</tbody>
</table>

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.

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 [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.

**Emotions.** Generally, Table 1 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 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 uni-modal 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 labelled, further research should be conducted.
CONCLUSION

In this article, we explored sub-symbolic 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 ten 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

Affective Computing and Sentiment Analysis

22. E. Marrese-Taylor, J. Balazs, and Y. Matsuo, “Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN,” *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, ACL, 2017, pp. 102–111.


**Alice Baird** received her MFA in Sound Art from Columbia University’s Computer Music Center and is currently a Ph.D. Fellow of the Center for digitization Bavaria (ZDB), supervised by Professor Prof. Björn Schuller at 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 interests include: health informatics, affective computing, and computational paralinguistics.

**Erik Cambria** is currently an Associate Professor with Nanyang Technological University, Singapore. His main research interests are AI and affective computing. He received the Ph.D. degree in computing science and mathematics through a joint programme 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** received his diploma, doctoral degree, habilitation, and Adjunct Teaching Professor in Machine Intelligence and Signal Processing from TUM in Munich, Germany. He is Full Professor of Artificial Intelligence and the Head of GLAM at Imperial College London/UK, Full Professor of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg, Germany amongst other Professorships and Affiliations. 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 AAC, and Senior Member of the ACM. He (co-)authored 1,000+ publications and was Editor in Chief of the IEEE Transactions on Affective Computing amongst further commitments.

**Lukas Stappen** received his MSc in Data Science with distinction from King’s College London in 2017. He then joined the group for Machine Learning in Health Informatics. Currently, he is a Ph.D. candidate at the Chair for Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany and a Ph.D. Fellow of the BMW Group in Munich. His research interests include affective computing, multimodal sentiment analysis, and multimodal representation learning with a core focus on ‘in-the-wild’ environments. Contact him at stappen@ieee.org.