YouTube Movie Reviews: In, Cross, and Open-domain Sentiment Analysis in an Audiovisual Context

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, Louis-Philippe Morency

Abstract—In this contribution we focus on the task of automatically analyzing a speaker's sentiment in on-line videos containing movie reviews. In addition to textual information, we consider adding audio features as typically used in speech-based emotion recognition as well as video features encoding valuable valence information conveyed by the speaker. We combine this multi-modal experimental setup with a detailed analysis of different methods for linguistic sentiment analysis by gradually increasing the level of domain-independence: First, we consider in-domain analysis by examining a cross-validation setup applied on a novel database named Multi-Modal Movie Opinion (ICT-MMMO) corpus. Next, we concentrate on cross-domain analysis by using a large corpus of written movie reviews for training. Finally, we explore the application of on-line knowledge sources for inferring the speaker's sentiment. Our experimental results indicate that training on written movie reviews is a promising alternative to exclusively using (spoken) in-domain data for building a system that analyses spoken movie review videos and that language-independent audiovisual analysis can compete with linguistic analysis.

Index Terms—sentiment analysis, affective computing, audiovisual pattern recognition, linguistic analysis

1 Introduction

Sentiment analysis, particularly the automatic analysis of written reviews in terms of positive or negative valence, has been extensively studied in the last decade. Many studies, e. g., [1], [2] classify reviews of products and services and report robust results for this application domain, such as 84 % accuracy for automobile reviews in [1]. In contrast, written movie reviews seem to be rather difficult to handle: In [1], 66 % accuracy of binary valence estimation are estimated for written movie reviews with the same method. One of the obvious challenges in classifying textual movie reviews is that sentiment words often relate to the elements of a movie rather than the reviewer's opinion. For instance, words one would usually associate with strongly negative valence, such as 'nightmare' or 'terrifying', could be used in a positive review of a horror movie.

As a first step towards more robust sentiment analysis in written movie reviews, usage of 'higher level' knowledge from on-line sources—including WordNet, ConceptNet, and General Inquirer—to better model the semantic relations between words in written movie reviews has been proposed in [3]. Furthermore, this study introduced a large (> 100 k instances) database of written movie reviews—the Metacritic database—which can be used for a robust data-based approach to written movie review classification: Contextual knowledge can be incorporated to a certain degree by relying on n-gram features, whose estimation usually requires large amounts of training data.

Arguably, besides linguistic cues, vocal expressions such as prosody and laughter, and facial expressions have to be taken into account for a holistic analysis of the speaker's sentiment. We expect that by fusing text-based sentiment classification with audio and video features, such as the ones often used in emotion recognition [4], the additional modalities can help classification in challenging cases such as those named above. Thus, building on [3], we now introduce multi-modal sentiment analysis in on-line review videos, which can be immediately applied in multimedia retrieval and tagging of large on-line video archives.

As a test database for this novel paradigm of sentiment analysis, we introduce a real-life collection of review videos obtained from the YouTube and ExpoTV platforms containing movie review videos by non-professional users. To create robust models, we further employ the large Metacritic database as training corpus, as well as knowledge from on-line sources including WordNet, ConceptNet, and General Inquirer; all of these are publicly available on the web. The crux is, however, that so far these resources have mostly been applied to written text—it is not clear how well they can cope with the peculiarities of spontaneous speech as often encountered in on-line review videos, including the prevalence of colloquialisms and malformed syntax (filled pauses, repetitions etc.) Thus, these resources will be compared to an approach

- *Martin Wöllmer, Felix Weninger, and Björn Schuller are with the Institute for Human-Machine Communication, Technische Universität München (TUM), Germany.*
- *Tobias Knaup is with Airbnb, California, USA.*
- *Congkai Sun, Kenji Sagae, and Louis-Philippe Morency are with the Institute for Creative Technology, University of Southern California, USA.*

relying on in-domain data consisting of transcriptions of spontaneous speech movie review videos.

2 Related Work

The work presented in this paper is closely related to two research fields: text-based sentiment analysis, which has been studied extensively in the field of computational linguistics, and audio-visual emotion recognition from the fields of speech processing and computer vision.

In text-based sentiment analysis, there is a growing body of work concerned with the automatic identification of sentiment in text, which often addresses online text, such as written reviews [1], [5], news articles, or blogs. While difficult problems such as cross-domain [6] or cross-language [7] portability have been addressed, not much has been done in terms of extending the applicability of sentiment analysis to other modalities, such as speech, gesture, or facial expressions. We are only aware of two exceptions. First, in the research reported in [8], speech and text are analyzed jointly for the purpose of subjectivity identification. This previous work, however, did not address other modalities such as visual cues, and did not address the problem of sentiment analysis. More recently, in a pre-study on 47 English review videos [9], it has been shown that visual and audio features can complement textual features for sentiment analysis.

For recent surveys of dimensional and categorical emotion recognition see [10]. In the related field of video retrieval, we have seen a new line of research addressing the multimodal fusion of language, acoustic features, and visual gestures, such as the VIRUS project that uses all three modalities to perform video retrieval [11].

In spite of these various publications dealing with text-based sentiment analysis and multimodal emotion recognition, a comprehensive study comparing in, cross, and open-domain sentiment analysis from acoustic, visual, and linguistic information obtained via automatic or manual transcription of on-line review videos does not exist so far, to the best of our knowledge. Hence, this article can be seen as a first attempt to evaluate these different aspects of sentiment analysis and to provide an impression of the corresponding accuracies for classification of a novel database of on-line videos containing spoken movie reviews.

3 Databases

3.1 ICT-MMMO: Multi-Modal Movie Opinion Database

With more than 10,000 videos being added every day, social media websites such as YouTube are well-suited for retrieving our dataset. People from all around the world post videos online and these videos are freely available. Also, social media websites contain the diversity, multi-modality and ambient noise characterizing real-world sentiment analysis.

We created a dataset, called ICT-MMMO (Multi-Modal Movie Opinion) Database from online social review videos that encompasses a strong diversity in how people express opinions about movies and includes a real-world variability in video recording quality (see http://multicomp.ict.usc.edu/). The dataset contains 370 multimodal review videos where one person is speaking directly at the camera, expressing their opinion and/or stating facts related to a specific movie.

Video Acquisition We collected 370 review videos from the social media websites YouTube and ExpoTV. The first part of our video collection started by search queries for movie review videos and opinions on the YouTube website, including optionally the name of recent movies as listed by imdb.com. An important challenge with movie review videos (and reviews in general) is that movies that originally received positive reviews have more chance of receiving follow-up reviews since more people will see these movies. In our first collection, out of 308 YouTube movie review videos, 228 review videos were annotated as positive while only 23 were annotated as neutral and 57 as negative.

We performed a second round of movie review video collection using the website ExpoTV.com which offers a forum for users to post review videos about movies, travel and products. Each review video is accompanied with a score from 1 to 5. We collected 78 movie review videos from ExpoTV, all of which with scores of 1 or 2. All these review videos were later annotated following the same sentiment annotation procedure used for the YouTube videos. This second video set was perceived as 62 negative, 14 neutral, and 2 positive review videos.

The final ICT-MMMO dataset includes all the 308 YouTube review videos and 62 negative movie review videos from ExpoTV, for a total of 370 movie review videos including 228 positive, 23 neutral and 119 negative video reviews. All speakers expressed themselves in English and the length of the review videos varies from 1-3 minutes.

Sentiment Annotation For this ICT-MMMO dataset, we are interested in the perceived sentiment expressed by the person being videotaped. To achieve this goal, all review videos were watched by coders who were instructed to assign one label per movie review video. We followed previous work on sentiment analysis [5] and used 5 sentiment labels, each associated with a numerical value: (1) strongly negative, (2) weakly negative, (3) neutral/ambivalent, (4) weakly positive and (5) strongly positive. All YouTube review videos were annotated by two coders while the ExpoTV review videos were annotated by only one coder given their original bias. It is important to note that we are not annotating the sentiment felt by the person watching the video. The annotation task is to associate a sentiment label that best summarizes the opinion expressed in the YouTube video. For the purpose of the experiments described in this paper, the sentiment annotations were averaged per review videos and categorized in two labels: negative ($\leq$3.5) and positive ($>$3.5). The threshold of 3.5 was chosen to obtain a comparable number of instances for

both classes and to separate positive from neutral and negative review videos as good as possible. We observe a very high inter-rater agreement for the YouTube review videos ($\kappa = 0.93$).

3.2 Metacritic Database

As an example for a large-scale on-line linguistic resource that can be used for data-based model training, we use the Metacritic database introduced in [3]. To the best of our knowledge, it still represents the largest corpus of written reviews used for sentiment classification. A total of 102 622 written reviews for 4 901 movies were downloaded from Metacritic (http://www.metacritic.com). Metacritic is a website that compiles written reviews for movies and other media mostly from on-line versions of newspapers and magazines; thus, most of the reviews are written by professional journalists. Written reviews in Metacritic are excerpts from the original texts consisting of usually one or two, mostly short, key sentences. Each written review in Metacritic is accompanied by a score which is mapped to the positive and negative valence classes following the schema proposed by Metacritic itself [3].

Comparing the Metacritic database with the ICT-MMMO corpus, it can be seen that they strongly differ by the length of the reviews (429 words on average for ICT-MMMO vs. 24 for Metacritic) and the language use: Many Metacritic reviews contain a very high level language including sophisticated metaphors and references, while the ICT-MMMO corpus is generally characterized by colloquial expressions and often malformed sentences.

4 On-Line Knowledge Sources

On-line knowledge sources (OKS) in natural language processing are databases of linguistic knowledge that are publicly available on the Internet. They contain information about words, concepts, or phrases, as well as connections among them. In [3], an approach is presented that uses three OKS to estimate valence of written movie reviews on the Metacritic database. General Inquirer is a lexical database that uses tags. Each entry consists of the term and a number of tags denoting the presence of a specific property in the term. WordNet is a database that organizes lexical concepts in terms of synonymy, meronymy or antonymy. ConceptNet is a database that contains a semantic network of commonsense knowledge. Concepts are interlinked by 26 different relations that encode the meaning of the connection between them. The idea of the algorithm used to infer sentiment scores via these OKS is to find the verbs and nouns which are 'closest' to affect related words as determined by General Inquirer. WordNet then serves to replace words unknown to General Inquirer by synonyms, and ConceptNet is used to 'filter out' expressions not relating to movies.

5 Multimodal Feature Extraction

5.1 Acoustic Features

For acoustic feature extraction we apply a large set of acoustic low-level descriptors (LLD) and derivatives of LLD combined with suited statistical functionals to capture speech dynamics within a turn (utterance between speech pauses). All features and functionals are computed using our on-line audio analysis toolkit openSMILE [12]. The audio feature set consists of 1 941 features and is identical to the feature set employed in [4]. It is composed of 25 energy and spectral related low-level descriptors x 42 functionals, 6 voicing related LLD x 32 functionals, 25 delta coefficients of the energy/spectral LLD x 23 functionals, 6 delta coefficients of the voicing related LLD x 19 functionals, and 10 voiced/unvoiced durational features.

In order to reduce the size of the resulting feature space, we apply a cyclic Correlation based Feature Subset Selection (CFS) [13] using the training set of each fold in our three-fold cross-validation experiments (see Section 7). For the three folds, this results in an automatic selection of 78, 74, and 71 acoustic features.

5.2 Video Features

The visual features are automatically extracted from the video sequences. Since only one person is present in each video clip and they are most of the time facing the camera, current technology for facial tracking [14] can efficiently be applied to our dataset.

As a first step, we used a commercial software, called OMRON's OKAO Vision System, which detects at each frame the face, extracts the facial features and extrapolates some basic facial expressions as well as eye gaze direction. The main facial expression being recognized is smile. This is a well-established technology that can be found in many digital cameras. For each frame, the vision software returns a smile intensity (0-100) and the gaze direction, using both horizontal and vertical angles expressed in degrees.

To complement these features, we processed all review videos using a 3D head pose tracker which is based on Generalized Adaptive View-based Appearance Model [15]. This method automatically acquires keyframes representing the head at different orientations and uses them to improve the tracker robustness and precision. At each frame, the tracker estimates the 3D position and orientation of the head. This information can be used to recognize absolute poses (e.g., head tilt or head down) as well as head gestures (e.g., head nods and head shakes). Both set of features were computed at the same rate as the original videos: 30 Hz.

Similar to our audio feature extraction method which produces one static feature vector per spoken utterance, we compute statistical functionals from the raw video feature vector sequences to obtain a fixed number of video descriptors for each turn. Thus, for every video feature stream, we compute the mean and standard deviation

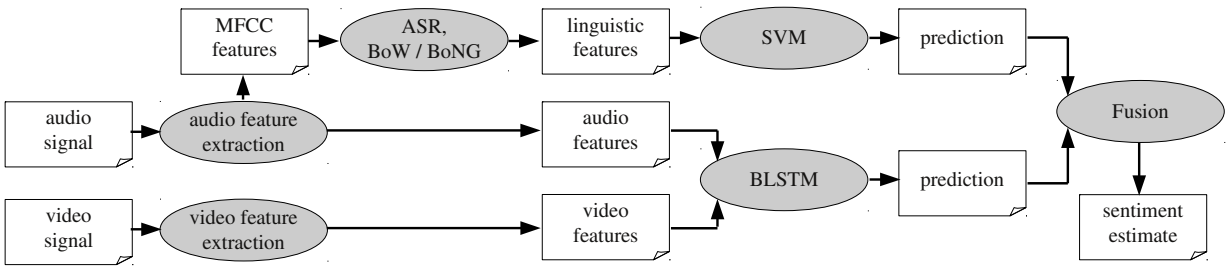


Fig. 1. System architecture for fusion of audiovisual and linguistic information (in and cross-domain analysis).

over a complete spoken utterance. This results in a video feature vector size of $2 \times 10 = 20$ features which is then reduced via CFS to a set of 6 features, on average.

5.3 Linguistic Features

As in [3], Bag-of-Words (BoW) and Bag-of-N-Gram (BoNG) features are used for data-based linguistic sentiment classification. The parametrization is taken from [3] and represents an optimal configuration on the Metacritic database, applying trigram features, Porter stemming, term frequency and inverse document frequency transformations, and document length normalization. To reduce the feature space, periodic pruning is applied and only the 1 000 features with the highest TF×IDF score in the training data are kept. Alternatively to generating linguistic features from the manual transcription of the ICT-MMMO database, we also apply an ASR system to obtain the transcriptions automatically. The ASR system is similar to the system used in [4] and was trained on the ICT-MMMO corpus in a cross-validation scheme (cf. Section 7).

6 Classification and Fusion

In order to model contextual information between successive utterances for sentiment analysis from audio and video features, we apply bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks. A detailed explanation of (B)LSTM networks can be found e. g. in [4].

For classification by linguistic features, linear Support Vector Machines (SVMs) are used. Figure 1 shows the overall system architecture we use for joint audiovisual and linguistic in and cross-domain sentiment analysis. Turnwise audio and video features are merged via early fusion and serve as input for the BLSTM network which in turn produces a sentiment prediction. An ASR system generates linguistic features from framewise MFCC features. The resulting BoW / BoNG features are classified via SVM so that a further prediction is produced. Note that the BLSTM network outputs a sentiment score for each spoken utterance whereas the SVM generates one prediction for each movie review video. Due to this asynchrony, late fusion is applied to infer the final sentiment estimate. The overall score generated by the BLSTM network is calculated by simply averaging the scores corresponding to the individual utterances. The final sentiment estimate is then computed as a weighted sum of the linguistic (weight 1.2) and the audiovisual (weight 0.8) score.

To integrate the scores of OKS into the above mentioned approach, they are mapped to the range $[0, 1]$ by means of logistic regression.

7 Experiments and Results

7.1 Experimental Setup

The knowledge-based approach (Section 4) is evaluated on the whole ICT-MMMO corpus. Our data-based approach (Sections 5 and 6) is applied in a within-domain setting as well as in a cross-domain setting. In the former, a three-fold cross-validation is performed on the ICT-MMMO corpus. The database is randomly split into three folds, yet we ensure that we have an equal number of different speakers in each fold and that the sets of speakers in the individual folds are disjoint. This reduces the danger of over-fitting to certain idiolects or interdependencies of speaker identity and sentiment polarity. Given the 343 transcribed ICT-MMMO movie review videos, the test sets of the three folds are of size 131, 99, and 113, respectively. In the cross-domain setting, the linguistic feature space and the model parameters are determined on the Metacritic corpus alone, and the ICT-MMMO corpus is used as a test set. By that, we can assess whether the features and models built from the Metacritic database of written, concise reviews generalize to the spontaneous speech review videos in the ICT-MMMO corpus. As evaluation measures, we rely on accuracy and weighted F1-measure, i. e., the average F1-measure of both classes weighted by their priors. In other words, the F1-measure used in our experiments is the F1-measure (harmonic mean of recall and precision) of the positive class weighted by the percentage of positive instances, plus the F1-measure of the negative class weighted by the percentage of negative instances. Furthermore, we

TABLE 1
Binary classification of sentiment polarity on the ICT-MMMO corpus by linguistic features: Accuracy (Acc.), (weighted) F1-measure, precision (Prec.) and recall (Rec.) for the positive (+) and negative (-) classes. Intra-corpus 3-fold cross-validation on the ICT-MMMO corpus or cross-corpus training on the Metacritic corpus. Linguistic features (Bag-of-Words, BoW, and Bag-of-N-Grams, BoNG) for the ICT-MMMO corpus generated either from manual transcription or from ASR.

Train on	Transcription	Features	Acc.	F1	Prec. (+)	Prec. (-)	Rec. (+)	Rec. (-)
ICT-MMMO	Manual	BoW	72.1	72.1	75.7	68.3	71.7	72.6
ICT-MMMO	Manual	BoNG	73.0	**73.0**	77.3	68.6	71.1	75.2
ICT-MMMO	ASR	BoW	63.7	63.7	68.5	59.1	61.5	66.2
ICT-MMMO	ASR	BoNG	58.4	57.9	66.9	53.3	46.5	72.6
Metacritic	Manual	BoW	67.4	67.1	78.2	60.7	55.6	81.5
Metacritic	Manual	BoNG	71.2	**71.3**	77.2	65.9	66.8	76.4
Metacritic	ASR	BoW	57.3	54.0	75.6	51.9	31.6	87.9
Metacritic	ASR	BoNG	61.0	60.9	68.0	55.8	53.5	70.1

consider the precision and recalls of both classes explicitly. Recall that when speaking of the ICT-MMMO corpus, we refer to review videos, while when speaking of the Metacritic database, we refer to written reviews.

7.2 Results of Linguistic Analysis

In Table 1, results of linguistic analysis by BoW and BoNG features are shown. Further, we compare features generated from the manual transcription and those inferred from ASR output. As expected, the overall best sentiment analysis accuracy and F1-measure (73.0 %) are achieved in the 'within-corpus' setting, i. e., 3-fold cross-validation. There, BoNG features slightly, yet not significantly outperform BoW features (72.1 %). Note that as a rule of thumb, performance differences of more than 6 % absolute are statistically significant ($p < .05$) according to a one-tailed z-test. However, it is notable that the performance in the cross-corpus setting, training on the Metacritic database, is observed only slightly below (up to 71.3 % F1 using BoNG features). In case of this extended training set, the BoNG features improve over BoW features by a larger margin than for the cross-validation, as would be expected. Concerning evaluation on features generated from ASR output, one has to accept a significant and consistent performance decrease of roughly 10 % absolute. The overall highest accuracy using ASR features (63.7 %) is achieved in cross-validation with BoW features; in the cross-corpus setting, 61.0 % are reached with BoNG features. With OKS, an accuracy of 59.6 % is estimated which is significantly above chance level, yet also significantly below the performance of in- or cross-domain analysis.

7.3 Results of Multi-Modal Fusion

In Table 2, we show the results of multi-modal fusion, i. e., we fuse the scores obtained by linguistic analysis with the BLSTM predictions obtained via audio and/or video features as depicted in Figure 1. Using audio features alone, an F1-measure of 63.8 % can be reached, which is remarkable considering that the audio-only system exclusively analyses the tone of the speaker's voice and does not consider any language information. Video features alone result in an F1-measure of 60.6 %, which is below the performance of audio features but still significantly above chance level. Applying combined audiovisual sentiment analysis, we get an F1-measure of 65.7 % which is higher than the results obtained via unimodal recognizers.

The performance gain obtained via fusion of linguistic and audiovisual information depends on the training scenario used for deriving the scores for linguistic analysis (in-domain, cross-domain, or open-domain) and on whether ASR is employed or not. For the in-domain experimental setup no noticeable performance difference can be seen when using different modality combinations together with linguistic analysis based on manual transcriptions. In case ASR is used, a slight improvement from 63.7 % to 65.0 % can be observed when adding audio information. The performance difference when using cross-domain analysis and ASR outputs is a bit more pronounced: Here, the F1-measure is increased from 60.9 % to 64.4 % when including audio features. The same holds for the open-domain case (improvement from 59.7 % to 64.2 % by adding audiovisual information). Overall, one can observe that audiovisual analysis only helps when linguistic analysis alone leads to low F1-measures, i. e., in the open-domain case or when linguistic analysis has to rely on error-prone ASR outputs.

The sensitivity of linguistic analysis to ASR errors is remarkable given recent studies in affective computing which show that emotion recognition tends to be robust with respect to errors made by the speech recognizer. This shows that in text which is more complex than the typically short emotionally-colored phrases used in studies on emotion recognition, textual accuracy seems to matter more.

8 Conclusion and Outlook

We introduced a system for sentiment analysis of on-line videos containing movie reviews by non-professional speakers as contained in a novel audiovisual database named ICT-MMMO corpus. The system applies bidirectional Long Short-Term Memory neural networks for

TABLE 2
Binary classification of sentiment polarity on the ICT-MMMO corpus by acoustic (A), video (V), and linguistic (L) features: Accuracy (Acc.), (weighted) F1-measure, precision (Prec.) and recall (Rec.) for the positive (+) and negative (-) classes. Intra-corpus 3-fold cross-validation on the ICT-MMMO corpus, cross-corpus training on the Metacritic corpus, or linguistic classification via on-line knowledge sources. Linguistic features (Bag-of-Words, BoW, and Bag-of-N-Grams, BoNG) for the ICT-MMMO corpus generated either from manual transcription or from ASR.

Train on (L)	Modalities	Transcription	Features (L)	Acc.	F1	Prec. (+)	Prec. (-)	Rec. (+)	Rec. (-)
-	A	-	-	64.4	63.8	64.7	64.0	75.8	51.0
-	V	-	-	61.2	60.6	62.2	59.5	72.6	47.8
-	AV	-	-	66.2	**65.7**	66.2	66.1	76.9	53.5
In-Domain: Linguistic classifier trained on ICT-MMMO corpus (test on ICT-MMMO corpus)									
ICT-MMMO	L	Manual	BoNG	73.0	73.0	77.3	68.6	71.1	75.2
ICT-MMMO	L+A	Manual	BoNG	72.3	72.4	76.3	68.2	71.0	73.9
ICT-MMMO	L+V	Manual	BoNG	73.2	**73.2**	77.7	68.8	71.0	75.8
ICT-MMMO	L+AV	Manual	BoNG	72.0	72.1	76.2	67.8	70.4	73.9
ICT-MMMO	L	ASR	BoW	63.7	63.7	68.5	59.1	61.5	66.2
ICT-MMMO	L+A	ASR	BoW	65.0	**65.0**	67.7	61.8	67.7	61.8
ICT-MMMO	L+V	ASR	BoW	61.5	61.6	65.3	57.5	61.8	61.1
ICT-MMMO	L+AV	ASR	BoW	62.1	62.2	65.7	58.2	62.9	61.1
Cross-Domain: Linguistic classifier trained on Metacritic corpus (test on ICT-MMMO corpus)									
Metacritic	L	Manual	BoNG	71.2	**71.3**	77.2	65.9	66.8	76.4
Metacritic	L+A	Manual	BoNG	71.1	71.1	72.8	69.1	74.7	66.9
Metacritic	L+V	Manual	BoNG	71.1	71.2	74.6	67.5	71.0	71.3
Metacritic	L+AV	Manual	BoNG	70.9	70.9	73.9	67.5	71.5	70.1
Metacritic	L	ASR	BoNG	61.0	60.9	68.0	55.8	53.5	70.1
Metacritic	L+A	ASR	BoNG	64.4	**64.4**	67.4	61.0	66.7	61.8
Metacritic	L+V	ASR	BoNG	63.0	63.0	67.5	58.6	61.3	65.0
Metacritic	L+AV	ASR	BoNG	63.9	63.9	67.8	59.8	63.4	64.3
Open-Domain: Linguistic classifier exploits on-line knowledge sources (test on ICT-MMMO corpus)									
-	L	Manual	-	59.6	59.7	64.0	55.2	58.8	60.5
-	L+A	Manual	-	64.7	63.8	64.2	65.8	79.0	47.8
-	L+V	Manual	-	64.7	63.6	63.9	66.4	80.1	46.5
-	L+AV	Manual	-	65.0	**64.2**	64.6	65.8	78.5	49.0

estimating the sentiment (positive vs. negative) conveyed in the review videos based on a set of audio and video features and on contextual information. For linguistic analysis, three different experimental setups are considered: *in-domain* analysis by training the linguistic analyzer on ICT-MMMO data in a cross-validation setting, *cross-domain* analysis by training on textual movie reviews found in the Metacritic database, and *open-domain* analysis by exploiting on-line knowledge sources.

The applied cross-corpus n-gram analysis based on the Metacritic database leads to remarkably high F1-measures of up to 71.3 % which are only slightly below within-corpus training (73.0 %). This implies that training on written reviews with scores retrieved automatically from the web is a promising method to classify spoken reviews, e. g., as contained in YouTube videos. The application of on-line knowledge sources cannot compete with n-gram models, however, the F1-measures obtained for linguistic analysis via on-line knowledge sources are significantly above chance level and can be improved by adding audiovisual information. Finally, we found that language-independent audiovisual analysis is almost as effective as in- and cross-domain linguistic analysis even though no textual information is used in this case.

Future work will concentrate on evaluations using larger databases, feature relevance analysis, and exploring methods for early or hybrid fusion of audiovisual and linguistic information for enhanced sentiment analysis. In contrast to the applied late fusion scheme, this would allow for an exploitation of complementary information during the classification process.

REFERENCES

[1] P. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews." in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, USA, 2002, pp. 417–424.

[2] E. Cambria, M. Grassi, A. Hussain, and C. Havasi, "Sentic computing for social media marketing," *Multimedia Tools and Applications*, vol. 59, no. 2, pp. 557–577, 2012.

[3] B. Schuller, J. Schenk, G. Rigoll, and T. Knaup, "'The Godfather' vs. 'Chaos': Comparing linguistic analysis based on online knowledge sources and bags-of-n-grams for movie review valence estimation," in *Proc. of ICDAR*, Barcelona, Spain, 2009, pp. 858–862.

[4] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, 2012.

[5] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain, 2004.

[6] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Association for Computational Linguistics*, 2007.

[7] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual subjectivity: Are more languages better?" in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China, August 2010, pp. 28–36. [Online]. Available: http://www.aclweb.org/anthology/C10-1004

[8] S. Raaijmakers, K. Truong, and T. Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, 2008, pp. 466–474.

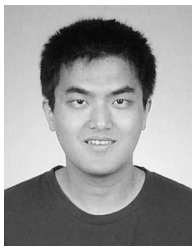
[9] L. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the International Conference on Multimodal Computing*, Alicante, Spain, 2011.
[10] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.
[11] P. Martins, T. Langlois, and T. Chambel, "Movieclouds: Content-based overviews and exploratory browsing of movies," in *Proceedings of the Academic MindTrek*, Tampere, Finland, 2011.
[12] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.
[13] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, University of Waikato, 1999.
[14] T. B. Dinh, N. Vo, and G. Medioni, "Context tracker: Exploring supporters and distracters in unconstrained environments," in *Proc. of CVPR*, Colorado Springs, USA, 2011.
[15] L. P. Morency, J. Whitehill, and J. Movellan, "Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation," in *Proc. of FG*, Amsterdam, The Netherlands, 2008.

Martin Wöllmer works as a research assistant at the Technische Universität München (TUM). He obtained his diploma in Electrical Engineering and Information Technology from TUM where his current research and teaching activity includes the subject areas of affective computing, pattern recognition and speech processing.

Felix Weninger (M'11) received his diploma in computer science from TUM. He is currently pursuing his Ph. D. degree in the Intelligent Audio Analysis Group at TUM's Institute for Human-Machine Communication. His research focuses on robust techniques for paralinguistic information retrieval from speech.

Tobias Knaup received his diploma in Electrical Engineering and Information Technology from TUM. He worked on a grant from the EC's CITPER project and spent a fellowship at AUB in Beirut/Lebanon. The focus of his research efforts is text-based Information Retrieval and Sentiment Analysis. At present he is a tech lead at Airbnb in San Francisco, California.

Björn Schuller received his diploma in 1999 and his doctoral degree for his study on Automatic Speech and Emotion Recognition in 2006, both in electrical engineering and information technology from TUM. He is tenured as Senior Lecturer in Pattern Recognition and Speech Processing heading the Intelligent Audio Analysis Group at TUM's Institute for Human-Machine Communication since 2006.



Congkai Sun did his bachelor at the Shanghai Jiaotong University, China in 2006. Then he started his Ph.D program at the University of Southern California (USC) in September 2009.

Kenji Sagae is a research assistant professor in the Computer Science department of the University of Southern California and a research scientist in the USC Institute for Creative Technologies. He received his Ph.D. from Carnegie Mellon University in 2006.

Louis-Philippe Morency received his Ph.D. from MIT Computer Science and Artificial Intelligence Laboratory. Dr. Louis-Philippe Morency is currently research assistant professor at the University of Southern California and research scientist at USC Institute for Creative Technologies where he leads the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab).