YouTube movie reviews: sentiment analysis in an audio-visual context

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

Angaben zur Veröffentlichung / Publication details:

Wollmer, Martin, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. "YouTube movie reviews: sentiment analysis in an audio-visual context." *IEEE Intelligent Systems* 28 (3): 46–53. https://doi.org/10.1109/mis.2013.34.

YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context

Martin Wöllmer, Felix Weninger, Tobias Knaup, and Björn Schuller, *Technische Universität München*

Congkai Sun, *Shanghai Jiaotong University*

Kenji Sagae and Louis-Philippe Morency, *University of Southern California*

This work focuses on automatically analyzing a speaker's sentiment in online videos containing movie reviews. In addition to textual information, this approach considers adding audio features typically used in speech-based emotion recognition, as well as video features encoding valuable valence information conveyed by the speaker.

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[1,2] classify reviews of products and services and report robust results for this application domain, such as 84 percent accuracy for automobile reviews in Peter Turney's work.[2] In contrast, written movie reviews seem to be rather difficult to handle: in Turney's work,[2] 66 percent 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 we 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, in previous work we proposed the use of "higher-level" knowledge from online sources—including WordNet, ConceptNet, and General Inquirer—to better model the semantic relations among words in written movie reviews.[3] Specifically, this work introduced a large database, called the *Metacritic database* (www.metacritic.com), with more than 100,000 instances of written movie reviews that 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 must 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

Related Work

Our work 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's a growing body of work concerned with the automatic identification of sentiment in text, which often addresses online text, such as written reviews,[1,2] news articles, or blogs. Although difficult problems such as cross-domain[3] or cross-language[4] 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're aware of only two exceptions. First, in Stephan Raaijmakers and his colleagues' research,[5] speech and text are analyzed jointly for the purpose of subjectivity identification. This work, however, didn't address other modalities such as visual cues, and it didn't address the problem of sentiment analysis. More recently, in a prestudy on 47 English review videos,[6] it has been shown that visual and audio features can complement textual features for sentiment analysis.

Zhihong Zeng and his colleagues provide a recent survey of dimensional and categorical emotion recognition.[7] In the related field of video retrieval, we've seen a new line of research addressing the multimodal fusion of language, acoustic features, and visual gestures, such as the Video Information Retrieval Using Subtitles (Virus) project that uses all three modalities to perform video retrieval.[8]

Despite 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 online review videos doesn't exist, 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 online videos containing spoken movie reviews.

References

1. P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proc. 40th Ann. Meeting of the Assoc. Computational Linguistics*, ACL, 2002, pp. 417–424.
2. B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proc. 42nd Meeting of the Assoc. Computational Linguistics*, ACL, 2004, pp. 271–278.
3. J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-Boxes, and Blenders: Domain Adaptation for Sentiment Classification," *Proc. Assoc. Computational Linguistics*, ACL, 2007, pp. 187–205.
4. C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual Subjectivity: Are More Languages Better?" *Proc. 23rd Int'l Conf. Computational Linguistics*, ACL, 2010, pp. 28–36.
5. S. Raaijmakers, K. Truong, and T. Wilson, "Multimodal Subjectivity Analysis of Multiparty Conversation," *Proc. Conf. on Empirical Methods in Natural Language Processing*, ACL, 2008, pp. 466–474.
6. L. Morency, R. Mihalcea, and P. Doshi, "Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web," *Proc. Int'l Conf. Multimodal Computing*, ACM, 2011, pp. 169–176.
7. Z. Zeng et al., "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, 2009, pp. 39–58.
8. P. Martins, T. Langlois, and T. Chambel, "Movieclouds: Content-Based Overviews and Exploratory Browsing of Movies," *Proc. Academic MindTrek*, ACM, 2011, pp. 133–140.

in challenging cases. Thus, building on this previous work,[3] we now introduce multimodal sentiment analysis in online review videos, which can be immediately applied in multimedia retrieval and tagging of large online video archives. (For others' work in this area, see the related sidebar.)

Databases

As a test database for this novel paradigm of sentiment analysis, we introduce a real-life collection of review videos obtained from YouTube and ExpoTV that contain movie review videos by nonprofessional users. To create robust models, we further employ the large Metacritic database as a training corpus, as well as knowledge from online sources such as 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 isn't clear how well they can cope with the peculiarities of spontaneous speech as often encountered in online review videos, including the prevalence of colloquialisms and malformed syntax (filled pauses, repetitions, and so on). Thus, these resources will be compared to an approach relying on in-domain data consisting of transcriptions of spontaneous speech movie review videos.

ICT-MMMO

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, multimodality, and ambient noise characterizing real-world sentiment analysis.

We therefore created a dataset, called the *Institute for Creative Technologies' Multi-Modal Movie Opinion* (ICT-MMMO) database, from online social review videos that encompass a strong diversity in how people express opinions about movies and include 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. We began collecting videos by using search queries for movie review videos and opinions on YouTube, sometimes including the names of recent movies. An important challenge with movie review videos (and reviews in general) is that movies that originally received positive reviews have a greater chance of receiving follow-up reviews because 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 ExpoTV website, which offers a forum for users to post review videos about movies, travel, and products. Each review video is accompanied by a score from 1–5. We collected 78 movie review videos from ExpoTV, all of which had 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 having 62 negative, 14 neutral, and 2 positive review videos.

The final ICT-MMMO dataset includes all 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 reviews. Every speaker spoke in English, and the length of the review videos varied from 1–3 minutes.

Sentiment annotation. For the ICT-MMMO dataset, we were interested in the perceived sentiment expressed by the person being videotaped. To achieve this goal, coders watched all of the review videos, and we instructed them to assign one label per movie review video. We followed previous work on sentiment analysis[5] and used five sentiment labels, each associated with a numerical value:

- strongly negative,
- weakly negative,
- neutral/ambivalent,
- weakly positive, and
- 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's important to note that we aren't 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 here, the sentiment annotations were averaged per review videos and categorized by two labels: negative (≤ 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 well as possible. We observed a high inter-rater agreement for the YouTube review videos ($\kappa = 0.93$).

Metacritic

As an example of a large-scale online linguistic resource that can be used for data-based model training, we used Metacritic.[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, a website that compiles written reviews for movies and other media mostly from online 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 and typically consist of one or two, mostly short, key sentences. Each written review in Metacritic is accompanied by a score that's mapped to positive and negative valence classes, following the schema proposed by Metacritic itself.[3]

Comparing the Metacritic database with the ICT-MMMO corpus, you can see that they strongly differ by the length of the reviews (429 words on average for ICT-MMMO versus 24 for Metacritic) and the language used. Many Metacritic reviews contain sophisticated metaphors and references, while the ICT-MMMO corpus is generally characterized by colloquial expressions and malformed sentences.

Online Knowledge Sources

Online 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. For example, in previous work, we used three OKS to estimate valence of written movie reviews on the Metacritic database: General Inquirer, WordNet, and ConceptNet.[3]

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 that are "closest" to affect-related words, as determined by General Inquirer. WordNet then replaces words unknown to General Inquirer with synonyms, and ConceptNet is used to "filter out" expressions unrelated to movies.

Multimodal Feature Extraction

Our system extracts acoustic, video, and linguistic features to aid in sentiment analysis.

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 (an utterance between speech pauses). All features and functionals are computed using our online audio analysis toolkit, openSMILE.[6] The audio feature set consists of 1,941 features and is identical to a feature set employed elsewhere.[4] It is composed of 25 energy and spectral-related low-level descriptors × 42 functionals, six voicing-related LLD × 32 functionals, 25 delta coefficients of the energy/spectral LLD × 23 functionals, six delta coefficients of the voicing-related LLD × 19 functionals, and 10 voiced/unvoiced durational features.

To reduce the size of the resulting feature space, we apply a cyclic correlation-based feature subset selection (CFS)[7] using the training set of each fold in our three-fold cross-validation experiments (which we detail later). For the three folds, this results in an automatic selection of 78, 74, and 71 acoustic features.

Video Features

The visual features are automatically extracted from the video sequences. Because only one person is present in each video clip, and most of the time they're facing the camera, current technology for facial tracking[8] can efficiently be applied to our dataset.

As a first step, we detect the face in every frame before we compute facial features and extrapolate a set of basic facial expressions and eye gaze direction using the commercial software Okao Vision. We focus on the smile as the most important facial expression and use a smile intensity from 0 to 100, which is returned by the software. In addition, gaze direction in the form of horizontal and vertical angles in degrees is applied.

To complement these features, we processed all review videos using a 3D head-pose tracker based on the Generalized Adaptive View-based Appearance Model.[9] This method automatically acquires keyframes representing the head at different orientations and uses them to improve tracker robustness and precision. At each frame, the tracker estimates the head's 3D position and orientation. This information can be used to recognize absolute poses (such as head tilt or head down) as well as head gestures (such as head nods and shakes). Both sets of features were computed at the same rate as the original videos: 30 Hz.

Similar to our audio feature-extraction method that produces one static feature vector per spoken utterance, we computed 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 computed the mean and standard deviation over a complete spoken utterance. This resulted in a video feature vector size of 2 × 10 = 20 features, which was then reduced via CFS to a set of six features, on average.

Linguistic Features

We use Bag-of-Words (BoW) and Bag-of-*N*-Gram (BoNG) features for data-based linguistic sentiment classification. The parameterization is taken from previous work,[3] and represents an optimal configuration on the Metacritic database, applying trigram features, Porter stemming, term frequency (TF), inverse document frequency (IDF) transformations, and document-length normalization. To reduce the feature space, periodic pruning is applied, and only the thousand 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 Automatic Speech Recognition (ASR) system to obtain the transcriptions automatically. The ASR system is similar to a system used elsewhere,[4] and was trained on the ICT-MMMO corpus in a cross-validation scheme.

Classification and Fusion

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 BLSTM networks can be found elsewhere.[4]

For classification by linguistic features, we use linear Support Vector Machines (SVMs). Figure 1 shows the overall system architecture we use for joint audio-visual 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 to produce further prediction. We should note that the BLSTM network outputs a sentiment score for each spoken utterance,

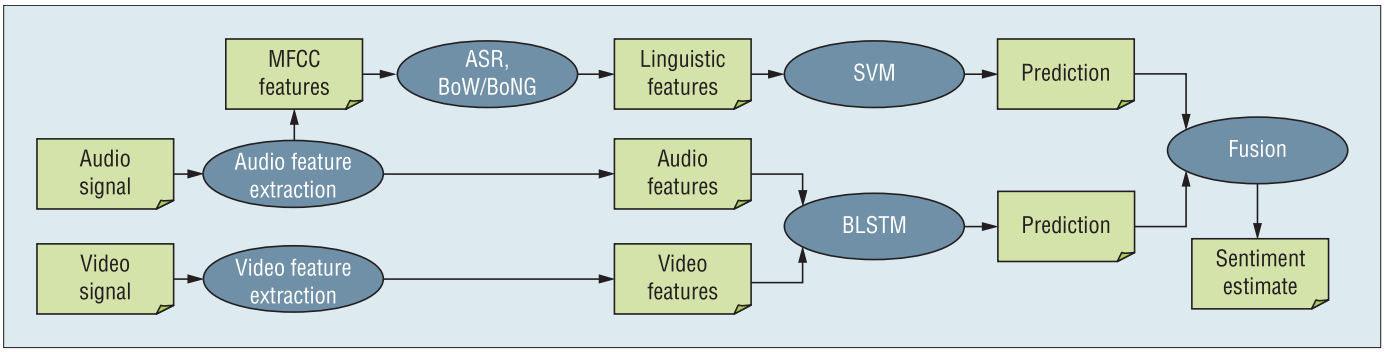

Figure 1. System architecture for fusion of audio-visual and linguistic information (for in- and cross-domain analysis). Turnwise audio and video features are merged via early fusion and serve as input for the bidirectional long short-term memory (BLSTM) network, which in turn produces a sentiment prediction.

whereas the SVM generates one prediction for each movie review video. Because of this asynchrony, we apply late fusion 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) scores.

To integrate the scores of OKS into the aforementioned approach, we map the scores to the range [0,1] by means of logistic regression.

Experiments and Results

After setting up the database and system, we tested our approach's performance.

Experimental Setup

We evaluated the knowledge-based approach on the whole ICT-MMMO corpus, applying our data-based approach in an in-domain setting as well as in a cross-domain setting. In the former, we performed a three-fold cross-validation on the ICT-MMMO corpus. We randomly split the database 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 a weighted F1-measure (the harmonic mean of recall and precision)—that is, 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 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. We also consider the precision and recalls of both classes explicitly. Finally, as we previously mentioned, the ICT-MMMO corpus refers to review videos, while the Metacritic database refers to written reviews.

Results of Linguistic Analysis

Table 1 shows the results of linguistic analysis by BoW and BoNG features. 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 percent) are achieved in the "within-corpus" setting, using a three-fold cross-validation. There, BoNG features slightly—yet not significantly—outperform BoW features (72.1 percent).

As a general rule, performance differences of more than 6 percent 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 percent F1 using BoNG features). In this instance, the BoNG features improve over BoW features by a larger margin than for the cross-validation. When evaluating features generated from ASR output, we must accept a significant and consistent performance decrease of roughly 10 percent absolute. The overall highest accuracy using ASR features (63.7 percent) is achieved in cross-validation with BoW features; in the cross-corpus setting, 61.0 percent are reached with BoNG features. With OKS, we estimate an accuracy of 59.6 percent, which is significantly above chance level, but significantly below the performance of in- or cross-domain analysis.

Results of Multimodal Fusion

Table 2 shows the results of multimodal fusion—that is, we fuse the scores obtained by linguistic analysis with the BLSTM predictions obtained

Table 1. Binary classification of sentiment polarity on the ICT-MMMO corpus by linguistic features.*

Training database	Transcription	Features	Accuracy	F1	Precision (+)	Precision (–)	Recall (+)	Recall (–)
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

*Intra-corpus three-fold cross-validation on the Institute for Creative Technologies' Multi-Modal Movie Opinion (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 Automatic Speech Recognition (ASR).

Table 2. Binary classification of sentiment polarity on the ICT-MMMO corpus by acoustic (A), video (V), and linguistic (L) features.*

Database used for training linguistic (L) classifier	Modalities	Transcription	Features (L)	Accuracy	F1	Precision (+)	Precision (–)	Recall (+)	Recall (–)
–	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 the 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 online 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

*Intra-corpus three-fold cross-validation on the ICT-MMMO corpus, cross-corpus training on the Metacritic corpus, or linguistic classification via online 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.

THE AUTHORS

Martin Wöllmer is a research assistant at the Technische Universität München. His current research interests include affective computing, pattern recognition, and speech processing. Wöllmer has a PhD in electrical engineering and information technology from TUM. Contact him at m.woellmer@gmx.de.

Felix Weninger is a PhD student in the Intelligent Audio Analysis Group at the Technische Universität München's Institute for Human-Machine Communication. His research focuses on robust techniques for paralinguistic information retrieval from speech. Weninger has a diploma in computer science from the Technische Universität München. Contact him at weninger@tum.de.

Tobias Knaup is a tech lead at Airbnb in San Francisco, California. His research interests include text-based information retrieval and sentiment analysis. Knaup has a diploma in electrical engineering and information technology from the Technical University of Munich. Contact him at tobi.knaup@gmail.com.

Björn Schuller leads the Machine Intelligence and Signal Processing Group at the Institute for Human-Machine Communication at the Technische Universität München. His research interests include machine learning, affective computing, and automatic speech recognition. Schuller has a PhD in electrical engineering and information technology from the Technische Universität München. Contact him at schuller@tum.de.

Congkai Sun is a PhD student at the University of Southern California. His research interests include computer science and pattern recognition. Sun has a BS in computer science from Shanghai Jiaotong University, China. Contact him at martin.sun.cn@gmail.com.

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. His research interests include computational linguistics. Sagae has a PhD in computer science from Carnegie Mellon University. Contact him at sagae@ict.usc.edu.

Louis-Philippe Morency is a research assistant professor in the Department of Computer Science at the University of Southern California and research scientist at the USC Institute for Creative Technologies, where he leads the Multimodal Communication and Machine Learning Laboratory. His research interests focus on the computational study of nonverbal social communication, a multidisciplinary research topic that includes multimodal interaction, computer vision, machine learning, social psychology, and artificial intelligence. Morency has a PhD in computer science and artificial intelligence from MIT. In 2008, *IEEE Intelligent Systems* selected him as one of the "10 to watch" for the future of AI research. Contact him at morency@ict.usc.edu.

via audio and/or video features as depicted in Figure 1. Using audio features alone, an F1-measure of 63.8 percent can be reached, which is remarkable considering that the audio-only system exclusively analyzes the tone of the speaker's voice and doesn't consider any language information. Video features alone result in an F1-measure of 60.6 percent, which is below the performance of audio features but still significantly above chance level. Applying combined audio-visual sentiment analysis, we get an F1-measure of 65.7 percent, which is higher than the results obtained via unimodal recognizers.

The performance gain obtained via fusion of linguistic and audio-visual 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. 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. When ASR is used, a slight improvement—from 63.7 to 65.0 percent—is 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 increases from 60.9 to 64.4 percent when including audio features. The same holds for the open-domain case, which increases from 59.7 to 64.2 percent with audio-visual information. Overall, then, audio-visual analysis helps only when linguistic analysis alone leads to low F1-measures—for instance, in the open-domain case or when linguistic analysis must 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 speech recognition errors. So, in text that is more complex than the short, emotionally colored phrases typically used in studies on emotion recognition, textual accuracy seems to matter more.

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

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

Acknowledgment

This material is based in part on work supported by US National Science Foundation award 1118018. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

1. E. Cambria et al., "Sentic Computing for Social Media Marketing," *Multimedia Tools and Applications*, vol. 59, no. 2, 2012, pp. 557–577.
2. P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proc. 40th Ann. Meeting of the Assoc. Computational Linguistics*, ACL, 2002, pp. 417–424.
3. B. Schuller et al., "'The Godfather' vs. 'Chaos': Comparing Linguistic Analysis Based on Online Knowledge Sources and Bags-of-N-Grams for Movie Review Valence Estimation," *Proc. Int'l Conf. Document Analysis and Recognition*, IEEE, 2009, pp. 858–862.
4. M. Wöllmer et al., "LSTM-Modeling of Continuous Emotions in an Audiovisual Affect Recognition Framework," *Image and Vision Computing*, vol. 31, no. 1, 2012, pp. 153–163.
5. B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proc. 42nd Meeting of the Assoc. Computational Linguistics*, ACL, 2004, pp. 271–278.
6. F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE—The Munich Versatile and Fast Open-Source Audio Feature Extractor," *Proc. ACM Multimedia*, ACM, 2010, pp. 1459–1462.
7. M. Hall, "Correlation-Based Feature Selection for Machine Learning," doctoral dissertation, Dept. of Computer Science, Univ. of Waikato, 1999.
8. T.B. Dinh, N. Vo, and G. Medioni, "Context Tracker: Exploring Supporters and Distracters in Unconstrained Environments," *Proc. Computer Vision and Pattern Recognition*, 2011, pp. 1177–1184.
9. L.P. Morency, J. Whitehill, and J. Movellan, "Generalized Adaptive View-Based Appearance Model: Integrated Framework for Monocular Head Pose Estimation," *Proc. Automatic Face and Gesture Recognition*, IEEE, 2008; doi:10.1109/AFGR.2008.4813429.