Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)

Towards Multimodal Emotion Recognition in German Speech Events in Cars using Transfer Learning

Deniz Cevher[1,2]*, **Sebastian Zepf**[1]* and **Roman Klinger**[2]
[1] Mercedes-Benz Research & Development, Daimler AG, Sindelfingen, Germany
[2] Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
{firstname.lastname}@daimler.com
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

The recognition of emotions by humans is a complex process which considers multiple interacting signals such as facial expressions and both prosody and semantic content of utterances. Commonly, research on automatic recognition of emotions is, with few exceptions, limited to one modality. We describe an in-car experiment for emotion recognition from speech interactions for three modalities: the audio signal of a spoken interaction, the visual signal of the driver's face, and the manually transcribed content of utterances of the driver. We use off-the-shelf tools for emotion detection in audio and face and compare that to a neural transfer learning approach for emotion recognition from text which utilizes existing resources from other domains. We see that transfer learning enables models based on out-of-domain corpora to perform well. This method contributes up to 10 percentage points in F_1, with up to 76 micro-average F_1 across the emotions joy, annoyance and insecurity. Our findings also indicate that off-the-shelf-tools analyzing face and audio are not ready yet for emotion detection in in-car speech interactions without further adjustments.

1 Introduction

Automatic emotion recognition is commonly understood as the task of assigning an emotion to a predefined instance, for example an utterance (as audio signal), an image (for instance with a depicted face), or a textual unit (e.g., a transcribed utterance, a sentence, or a Tweet). The set of emotions is often following the original definition by Ekman (1992), which includes anger, fear, disgust, sadness, joy, and surprise, or the extension by Plutchik (1980) who adds trust and anticipation.

Most work in emotion detection is limited to one modality. Exceptions include Busso et al. (2004) and Sebe et al. (2005), who investigate multimodal approaches combining speech with facial information. Emotion recognition in speech can utilize semantic features as well (Anagnostopoulos et al., 2015). Note that the term "multimodal" is also used beyond the combination of vision, audio, and text. For example, Soleymani et al. (2012) use it to refer to the combination of electroencephalogram, pupillary response and gaze distance.

In this paper, we deal with the specific situation of car environments as a testbed for multimodal emotion recognition. This is an interesting environment since it is, to some degree, a controlled environment: Dialogue partners are limited in movement, the degrees of freedom for occurring events are limited, and several sensors which are useful for emotion recognition are already integrated in this setting. More specifically, we focus on emotion recognition from speech events in a dialogue with a human partner and with an intelligent agent.

Also from the application point of view, the domain is a relevant choice: Past research has shown that emotional intelligence is beneficial for human computer interaction. Properly processing emotions in interactions increases the engagement of users and can improve performance when a specific task is to be fulfilled (Klein et al., 2002; Coplan and Goldie, 2011; Partala and Surakka, 2004; Pantic et al., 2005). This is mostly based on the aspect that machines communicating with humans appear to be more trustworthy when they show empathy and are perceived as being natural (Partala and Surakka, 2004; Brave et al., 2005; Pantic et al., 2005).

Virtual agents play an increasingly important role in the automotive context and the speech modality is increasingly being used in cars due to its potential to limit distraction. It has been shown

*The first two authors contributed equally.

that adapting the in-car speech interaction system according to the drivers' emotional state can help to enhance security, performance as well as the overall driving experience (Nass et al., 2005; Harris and Nass, 2011).

With this paper, we investigate how each of the three considered modalitites, namely facial expressions, utterances of a driver as an audio signal, and transcribed text contributes to the task of emotion recognition in in-car speech interactions. We focus on the five emotions of *joy*, *insecurity*, *annoyance*, *relaxation*, and *boredom* since terms corresponding to so-called fundamental emotions like *fear* have been shown to be associated to too strong emotional states than being appropriate for the in-car context (Dittrich and Zepf, 2019). Our first contribution is the description of the experimental setup for our data collection. Aiming to provoke specific emotions with situations which can occur in real-world driving scenarios and to induce speech interactions, the study was conducted in a driving simulator. Based on the collected data, we provide baseline predictions with off-the-shelf tools for face and speech emotion recognition and compare them to a neural network-based approach for emotion recognition from text. Our second contribution is the introduction of transfer learning to adapt models trained on established out-of-domain corpora to our use case. We work on German language, therefore the transfer consists of a domain and a language transfer.

2 Related Work

2.1 Facial Expressions

A common approach to encode emotions for facial expressions is the facial action coding system FACS (Ekman and Friesen, 1978; Sujono and Gunawan, 2015; Lien et al., 1998). As the reliability and reproducibility of findings with this method have been critically discussed (Mesman et al., 2012), the trend has increasingly shifted to perform the recognition directly on images and videos, especially with deep learning. For instance, Jung et al. (2015) developed a model which considers temporal geometry features and temporal appearance features from image sequences. Kim et al. (2016) propose an ensemble of convolutional neural networks which outperforms isolated networks.

In the automotive domain, FACS is still popular. Ma et al. (2017) use support vector machines to distinguish *happy*, *bothered*, *confused*, and *concentrated* based on data from a natural driving environment. They found that *bothered* and *confused* are difficult to distinguish, while *happy* and *concentrated* are well identified. Aiming to reduce computational cost, Tews et al. (2011) apply a simple feature extraction using four dots in the face defining three facial areas. They analyze the variance of the three facial areas for the recognition of *happy*, *anger* and *neutral*. Ihme et al. (2018) aim at detecting *frustration* in a simulator environment. They induce the emotion with specific scenarios and a demanding secondary task and are able to associate specific face movements according to FACS. Paschero et al. (2012) use OpenCV (https://opencv.org/) to detect the eyes and the mouth region and track facial movements. They simulate different lightning conditions and apply a multilayer perceptron for the classification task of Ekman's set of fundamental emotions.

Overall, we found that studies using facial features usually focus on continuous driver monitoring, often in driver-only scenarios. In contrast, our work investigates the potential of emotion recognition during speech interactions.

2.2 Acoustic

Past research on emotion recognition from acoustics mainly concentrates on either feature selection or the development of appropriate classifiers. Rao et al. (2013) as well as Ververidis et al. (2004) compare local and global features in support vector machines. Next to such discriminative approaches, hidden Markov models are well-studied, however, there is no agreement on which feature-based classifier is most suitable (El Ayadi et al., 2011). Similar to the facial expression modality, recent efforts on applying deep learning have been increased for acoustic speech processing. For instance, Lee and Tashev (2015) use a recurrent neural network and Palaz et al. (2015) apply a convolutional neural network to the raw speech signal. Neumann and Vu (2017) as well as Trigeorgis et al. (2016) analyze the importance of features in the context of deep learning-based emotion recognition.

In the automotive sector, Boril et al. (2011) approach the detection of negative emotional states within interactions between driver and co-driver as well as in calls of the driver towards the automated spoken dialogue system. Using real-world driving data, they find that the combination of acoustic features and their respective Gaussian mixture model

scores performs best. Schuller et al. (2006) collects 2,000 dialog turns directed towards an automotive user interface and investigate the classification of *anger*, *confusion*, and *neutral*. They show that automatic feature generation and feature selection boost the performance of an SVM-based classifier. Further, they analyze the performance under systematically added noise and develop methods to mitigate negative effects. For more details, we refer the reader to the survey by Schuller (2018). In this work, we explore the straight-forward application of domain independent software to an in-car scenario without domain-specific adaptations.

2.3 Text

Previous work on emotion analysis in natural language processing focuses either on resource creation or on emotion classification for a specific task and domain. On the side of resource creation, the early and influential work of Pennebaker et al. (2015) is a dictionary of words being associated with different psychologically relevant categories, including a subset of emotions. Another popular resource is the NRC dictionary by Mohammad and Turney (2012). It contains more than 10000 words for a set of discrete emotion classes. Other resources include WordNet Affect (Strapparava and Valitutti, 2004) which distinguishes particular word classes. Further, annotated corpora have been created for a set of different domains, for instance fairy tales (Alm et al., 2005), Blogs (Aman and Szpakowicz, 2007), Twitter (Mohammad et al., 2017; Schuff et al., 2017; Mohammad, 2012; Mohammad and Bravo-Marquez, 2017a; Klinger et al., 2018), Facebook (Preoţiuc-Pietro et al., 2016), news headlines (Strapparava and Mihalcea, 2007), dialogues (Li et al., 2017), literature (Kim et al., 2017), or self reports on emotion events (Scherer, 1997) (see (Bostan and Klinger, 2018) for an overview).

To automatically assign emotions to textual units, the application of dictionaries has been a popular approach and still is, particularly in domains without annotated corpora. Another approach to overcome the lack of huge amounts of annotated training data in a particular domain or for a specific topic is to exploit distant supervision: use the signal of occurrences of emoticons or specific hashtags or words to automatically label the data. This is sometimes referred to as self-labeling (Klinger et al., 2018; Pool and Nissim, 2016; Felbo et al., 2017; Wang et al., 2012).

Figure 1: The setup of the driving simulator.

A variety of classification approaches have been tested, including SNoW (Alm et al., 2005), support vector machines (Aman and Szpakowicz, 2007), maximum entropy classification, long short-term memory network, and convolutional neural network models (Schuff et al., 2017, *i.a.*). More recently, the state of the art is the use of transfer learning from noisy annotations to more specific predictions (Felbo et al., 2017). Still, it has been shown that transferring from one domain to another is challenging, as the way emotions are expressed varies between areas (Bostan and Klinger, 2018). The approach by Felbo et al. (2017) is different to our work as they use a huge noisy data set for pretraining the model while we use small high quality data sets instead.

Recently, the state of the art has also been pushed forward with a set of shared tasks, in which the participants with top results mostly exploit deep learning methods for prediction based on pretrained structures like embeddings or language models (Klinger et al., 2018; Mohammad et al., 2018; Mohammad and Bravo-Marquez, 2017a).

Our work follows this approach and builds up on embeddings with deep learning. Furthermore, we approach the application and adaption of text-based classifiers to the automotive domain with transfer learning.

3 Data set Collection

The first contribution of this paper is the construction of the AMMER data set which we describe in the following. We focus on the drivers' interactions with both a virtual agent as well as a co-driver. To collect the data in a safe and controlled environment and to be able to consider a variety of predefined driving situations, the study was conducted in a driving simulator.

Type	Example
D–A, beginning	Wie geht es dir gerade und wie sind deine Gedanken zur bevorstehenden Fahrt? *How are you doing right now? What are your thoughts about the upcoming drive?*
D–A, reaching destination	Bei über 50 Teilnehmern hast du die zweitschnellste Zeit erreicht. Was glaubst du? Wie hast du es geschafft so schnell zu sein? *Among more than 50 participants you achieved the second best result. What do you think? How did you manage to achieve that?*
D–A, after driving	Du hast im letzten Streckenabschnitt ein paar Mal stark gebremst. Was ist da passiert? *In the last section, you slowed down multiple times. What happened?*
D–Co, low-demand section	Erinnern Sie sich an Ihren letzten Urlaub. Bitte beschreiben Sie, wie dieser Urlaub für Sie war? *Remember your last vacation. Please describe how it was.*

Table 1: Examples for triggered interactions with translations to English. (D: Driver, A: Agent, Co: Co-Driver)

3.1 Study Setup and Design

The study environment consists of a fixed-base driving simulator running Vires's VTD (Virtual Test Drive, v2.2.0) simulation software (https://vires.com/vtd-vires-virtual-test-drive/). The vehicle has an automatic transmission, a steering wheel and gas and brake pedals. We collect data from video, speech and biosignals (Empatica E4 to record heart rate, electrodermal activity, skin temperature, not further used in this paper) and questionnaires. Two RGB cameras are fixed in the vehicle to capture the drivers face, one at the sun shield above the drivers seat and one in the middle of the dashboard. A microphone is placed on the center console. One experimenter sits next to the driver, the other behind the simulator. The virtual agent accompanying the drive is realized as Wizard-of-Oz prototype which enables the experimenter to manually trigger prerecorded voice samples playing trough the in-car speakers and to bring new content to the center screen. Figure 1 shows the driving simulator.

The experimental setting is comparable to an everyday driving task. Participants are told that the goal of the study is to evaluate and to improve an intelligent driving assistant. To increase the probability of emotions to arise, participants are instructed to reach the destination of the route as fast as possible while following traffic rules and speed limits. They are informed that the time needed for the task would be compared to other participants. The route comprises highways, rural roads, and city streets. A navigation system with voice commands and information on the screen keeps the participants on the predefined track.

To trigger emotion changes in the participant, we use the following events: (i) a car on the right lane cutting off to the left lane when participants try to overtake followed by trucks blocking both lanes with a slow overtaking maneuver (ii) a skateboarder who appears unexpectedly on the street and (iii) participants are praised for reaching the destination unexpectedly quickly in comparison to previous participants.

Based on these events, we trigger three interactions (Table 1 provides examples) with the intelligent agent (*Driver-Agent Interactions, D–A*). Pretending to be aware of the current situation, *e. g.*, to recognize unusual driving behavior such as strong braking, the agent asks the driver to explain his subjective perception of these events in detail. Additionally, we trigger two more interactions with the intelligent agent at the beginning and at the end of the drive, where participants are asked to describe their mood and thoughts regarding the (upcoming) drive. This results in five interactions between the driver and the virtual agent.

Furthermore, the co-driver asks three different questions during sessions with light traffic and low cognitive demand (*Driver-Co-Driver Interactions, D–Co*). These questions are more general and non-traffic-related and aim at triggering the participants' memory and fantasy. Participants are asked to describe their last vacation, their dream house and their idea of the perfect job. In sum, there are eight interactions per participant (5 D–A, 3 D–Co).

3.2 Procedure

At the beginning of the study, participants were welcomed and the upcoming study procedure was explained. Subsequently, participants signed a consent form and completed a questionnaire to provide demographic information. After that, the co-driving experimenter started with the instruction

E	IT	Example
J	A	Ich glaube, weil ich ziemlich schnell auf Situationen reagieren kann, weil ich eine ziemlich gute Reaktion habe. Und ich würde auch behaupten, dass ich relativ vorausschauend fahre, weil ich schon einiges an Fahrerfahrung mitbringe. *I think because I can respond to situations very quickly because my reaction is very good. And I would say that I drive foresightful because I have a lot of driving experience.*
J	C	Letzter Urlaub war im September 2018. Singapur und Bali. War sehr schön. Erholung, andere Kultur, andere Länder. War sehr gut und ist zu wiederholen. *Last vacation was in September 2018. Singapore and Bali. It was beautiful. Recreation, different culture, different countries. It was very good and needs repetition.*
A	A	Zwei bis drei Mal Fahrzeuge, die Kolonne fuhren. Und das letzte Fahrzeug hat, für mein Gefühl, sehr ruckartig und mit wenig nach hinten zu schauen, die Spur gewechselt und mich dazu gezwungen, dann doch noch meine Geschwindigkeit zu reduzieren. *Two or three times vehicles were driving behind each other. The last vehicle cut off my lane, in my opinion very quickly and without looking back and forced me to slow down.*
A	C	Mir geht es nicht besonders gut. Die Fahrt war sehr stressig. Ich schwitze ziemlich. *I'm not feeling well. The ride was stressful. I am sweating.*
I	A	Letzter Urlaub war nicht so gut für mich. Obwohl. Naja doch. Der letzte war schon wieder gut. Das war im Sommer. Da war es nämlich so abartig warm dieses Jahr. Und wir haben bei uns daheim. Also ich komme ja vom Land. Wir haben bei uns daheim auf dem Land unseren Wohnwagen ausgebaut. *Last vacation was not so good for me. Although. Well, yes. The last one was good. It was in summer. It was very warm this year. And we have at home. I come from the countryside. We have furnished our mobile home.*
I	C	Ein Mensch ist über die Straße gelaufen und ich habe ihn zuerst nicht gesehen. *A human crossed the street and I haven't seen him in the first moment.*
B	A	Ich habe mich immer an die Richtgeschwindigkeit gehalten. Und ja. Ich weiß auch nicht. *I always followed the recommended velocity. And, well. I don't know.*
B	C	Ja. Nicht viel arbeiten und viel Geld verdienen. *Yes. Not working much and earning a lot of money.*
R	A	Mir geht es gut und ich bin gespannt auf die Fahrt. Ich denke, es macht Spaß. *I am fine and I am looking forward to the ride. I think it will be fun.*
R	C	Ja, ich erinnere mich an den letzten Urlaub und der war schön, war erholsam und war warm. *Yes, I remember the last vacation. It was nice, recreative and warm.*
N	A	Es sind Autos von der rechten Spur auf meine Spur gezogen, welche davor deutlich langsamer waren. *Cars were changing into my lane, which were slower before.*
N	C	Ein Haus, das relativ alleine für sich steht. Am besten am Meer und mit einem grünen Garten. Und ja. Viel Platz für sich. *A house with space around. In the best case at the sea and with a green garden. And yes. A lot of space for us.*

Table 2: Examples from the collected data set (with translation to English). E: Emotion, IT: interaction type with agent (A) and with Codriver (C). J: Joy, A: Annoyance, I: Insecurity, B: Boredom, R: Relaxation, N: No emotion.

in the simulator which was followed by a familiarization drive consisting of highway and city driving and covering different driving maneuvers such as tight corners, lane changing and strong braking. Subsequently, participants started with the main driving task. The drive had a duration of 20 minutes containing the eight previously mentioned speech interactions. After the completion of the drive, the actual goal of improving automatic emotional recognition was revealed and a standard emotional intelligence questionnaire, namely the TEIQue-SF (Cooper and Petrides, 2010), was handed to the participants. Finally, a retrospective interview was conducted, in which participants were played recordings of their in-car interactions and asked to give discrete (annoyance, insecurity, joy, relaxation, boredom, none, following (Dittrich and Zepf, 2019)) was well as dimensional (valence, arousal, dominance (Posner et al., 2005) on a 11-point scale) emotion ratings for the interactions and the according situations. We only use the discrete class annotations in this paper.

3.3 Data Analysis

Overall, 36 participants aged 18 to 64 years (μ=28.89, σ=12.58) completed the experiment. This leads to 288 interactions, 180 between driver and the agent and 108 between driver and co-driver. The emotion self-ratings from the participants yielded 90 utterances labeled with *joy*, 26 with *annoyance*, 49 with *insecurity*, 9 with *boredom*, 111 with *relaxation* and 3 with *no emotion*. One example interaction per interaction type and emotion is shown in Table 2. For further experiments, we only use joy, annoyance/anger, and insecurity/fear due to the small sample size for boredom and no emotion and under the assumption that relaxation brings little expressivity.

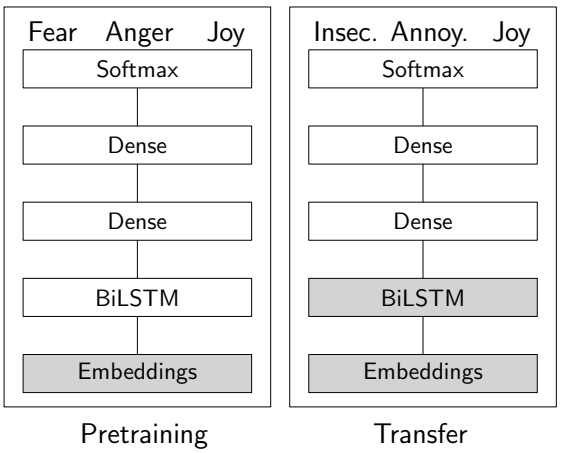

Figure 2: Model for Transfer Learning from Text. Grey boxes contain frozen parameters in the corresponding learning step.

4 Methods

4.1 Emotion Recognition from Facial Expressions

We preprocess the visual data by extracting the sequence of images for each interaction from the point where the agent's or the co-driver's question was completely uttered until the driver's response stops. The average length is 16.3 seconds, with the minimum at 2.2s and the maximum at 54.7s. We apply an off-the-shelf tool for emotion recognition (the manufacturer cannot be disclosed due to licensing restrictions). It delivers frame-by-frame scores ($\in [0;100]$) for discrete emotional states of *joy*, *anger* and *fear*. While *joy* corresponds directly to our annotation, we map *anger* to our label *annoyance* and *fear* to our label *insecurity*. The maximal average score across all frames constitutes the overall classification for the video sequence. Frames where the software is not able to detect the face are ignored.

4.2 Emotion Recognition from Audio Signal

We extract the audio signal for the same sequence as described for facial expressions and apply an off-the-shelf tool for emotion recognition. The software delivers single classification scores for a set of 24 discrete emotions for the entire utterance. We consider the outputs for the states of joy, anger, and fear, mapping analogously to our classes as for facial expressions. Low-confidence predictions are interpreted as "no emotion". We accept the emotion with the highest score as the discrete prediction otherwise.

4.3 Emotion Recognition from Transcribed Utterances

For the emotion recognition from text, we manually transcribe all utterances of our AMMER study. To exploit existing and available data sets which are larger than the AMMER data set, we develop a transfer learning approach. We use a neural network with an embedding layer (frozen weights, pretrained on Common Crawl and Wikipedia (Grave et al., 2018)), a bidirectional LSTM (Schuster and Paliwal, 1997), and two dense layers followed by a soft max output layer. This setup is inspired by (Andryushechkin et al., 2017). We use a dropout rate of 0.3 in all layers and optimize with Adam (Kingma and Ba, 2015) with a learning rate of 10^{-5} (These parameters are the same for all further experiments). We build on top of the Keras library with the TensorFlow backend. We consider this setup our *baseline model*.

We train models on a variety of corpora, namely the common format published by (Bostan and Klinger, 2018) of the *FigureEight* (formally known as Crowdflower) data set of social media, the *ISEAR* data (Scherer and Wallbott, 1994) (self-reported emotional events), and, the Twitter Emotion Corpus (TEC, weakly annotated Tweets with #anger, #disgust, #fear, #happy, #sadness, and #surprise, Mohammad (2012)). From all corpora, we use instances with labels fear, anger, or joy. These corpora are English, however, we do predictions on German utterances. Therefore, each corpus is preprocessed to German with Google Translate[1]. We remove URLs, user tags ("@Username"), punctuation and hash signs. The distributions of the data sets are shown in Table 3.

To adapt models trained on these data, we apply *transfer learning* as follows: The model is first trained until convergence on one out-of-domain corpus (only on classes fear, joy, anger for compatibility reasons). Then, the parameters of the bi-LSTM layer are frozen and the remaining layers are further trained on AMMER. This procedure is illustrated in Figure 2

5 Results

5.1 Facial Expressions and Audio

Table 4 shows the confusion matrices for facial and audio emotion recognition on our complete AMMER data set and Table 5 shows the re-

[1] http://translate.google.com, performed on January 4, 2019

Data set	Fear	Anger	Joy	Total
Figure8	8,419	1,419	9,179	19,017
EmoInt	2,252	1,701	1,616	5,569
ISEAR	1,095	1,096	1,094	3,285
TEC	2,782	1,534	8,132	12,448
AMMER	49	26	90	165

Table 3: Class distribution of the used data sets for the considered emotional states (Figure8 (Figure Eight, 2016), EmoInt (Mohammad and Bravo-Marquez, 2017b), ISEAR, (Scherer, 1997), TEC (Mohammad, 2012), AMMER (this paper)).

	Vision				
	Fear	Anger	Joy		Total
Insecurity	11	17	21		49
Annoyance	10	7	9		26
Joy	24	27	39		90
Total	45	51	69		165
	Audio				
	Fear	Anger	Joy	No	Total
Insecurity	17	14	1	17	49
Annoyance	12	7	0	7	26
Joy	27	26	4	33	90
Total	56	47	5	57	165
	Transfer Learning Text				
	Fear	Anger	Joy	No	Total
Insecurity	33	0	16		49
Annoyance	7	4	15		26
Joy	1	1	88		90
Total	41	5	119		165

Table 4: Confusion Matrix for Face Classification and Audio Classification (on full AMMER data) and for transfer learning from text (training set of EmoInt and test set of AMMER). Insecurity, annoyance and joy are the gold labels. Fear, anger and joy are predictions.

sults per class for each method, including facial and audio data and micro and macro averages. The classification from facial expressions yields a macro-averaged F_1 score of 33 % across the three emotions joy, insecurity, and annoyance (P=0.31, R=0.35). While the classification results for joy are promising (R=43 %, P=57 %), the distinction of insecurity and annoyance from the other classes appears to be more challenging.

Regarding the audio signal, we observe a macro F_1 score of 29 % (P=42 %, R=22 %). There is a bias towards negative emotions, which results in a small number of detected joy predictions (R=4 %). Insecurity and annoyance are frequently confused.

5.2 Text from Transcribed Utterances

The experimental setting for the evaluation of emotion recognition from text is as follows: We evaluate the BiLSTM model in three different experiments: (1) in-domain, (2) out-of-domain and (3) transfer learning. For all experiments we train on the classes *anger/annoyance*, *fear/insecurity* and *joy*. Table 6 shows all results for the comparison of these experimental settings.

5.2.1 Experiment 1: In-Domain application

We first set a baseline by validating our models on established corpora. We train the baseline model on 60 % of each data set listed in Table 3 and evaluate that model with 40 % of the data from the same domain (results shown in the column "In-Domain" in Table 6). Excluding AMMER, we achieve an average micro F_1 of 68 %, with best results of F_1=73 % on TEC. The model trained on our AMMER corpus achieves an F1 score of 57%. This is most probably due to the small size of this data set and the class bias towards *joy*, which makes up more than half of the data set. These results are mostly in line with Bostan and Klinger (2018).

5.2.2 Experiment 2: Simple Out-Of-Domain application

Now we analyze how well the models trained in Experiment 1 perform when applied to our data set. The results are shown in column "Simple" in Table 6. We observe a clear drop in performance, with

	Vision			Audio			Text (TL)		
	P	R	F_1	P	R	F_1	P	R	F_1
Insecurity	24	22	23	31	35	33	80	67	73
Annoyance	14	39	21	15	27	19	80	15	26
Joy	57	43	49	80	4	8	74	98	84
Macro-avg	32	35	33	42	22	29	78	60	68
Micro-avg	34	34	34	26	17	21	76	76	76

Table 5: Performance for classification from vision, audio, and transfer learning from text (training set of EmoInt).

		Out-of-domain		
Train Corpus	In-Domain	Simple	Joint C.	Transfer L.
Figure8	66	55	59	76
EmoInt	62	48	56	76
TEC	73	55	58	76
ISEAR	70	35	59	72
AMMER	57	—	—	—

Table 6: Results in micro F_1 for Experiment 1 (in-domain), Experiment 2 and 3 (out-of-domain with and without transfer learning).

an average of F_1=48 %. The best performing model is again the one trained on TEC, en par with the one trained on the Figure8 data. The model trained on ISEAR performs second best in Experiment 1, it performs worst in Experiment 2.

5.2.3 Experiment 3: Transfer Learning application

To adapt models trained on previously existing data sets to our particular application, the AMMER corpus, we apply transfer learning. Here, we perform leave-one-out cross validation. As pre-trained models we use each model from Experiment 1 and further optimize with the training subset of each crossvalidation iteration of AMMER. The results are shown in the column "Transfer L." in Table 6. The confusion matrix is also depicted in Table 4.

With this procedure we achieve an average performance of F_1=75 %, being better than the results from the in-domain Experiment 1. The best performance of F_1=76 % is achieved with the model pre-trained on each data set, except for ISEAR. All transfer learning models clearly outperform their simple out-of-domain counterpart.

To ensure that this performance increase is not only due to the larger data set, we compare these results to training the model without transfer on a corpus consisting of each corpus together with AMMER (again, in leave-one-out crossvalidation). These results are depicted in column "Joint C.". Thus, both settings, "transfer learning" and "joint corpus" have access to the same information.

The results show an increase in performance in contrast to not using AMMER for training, however, the transfer approach based on partial retraining the model shows a clear improvement for all models (by 7pp for Figure8, 10pp for EmoInt, 8pp for TEC, 13pp for ISEAR) compared to the "Joint" setup.

6 Summary & Future Work

We described the creation of the multimodal AMMER data with emotional speech interactions between a driver and both a virtual agent and a co-driver. We analyzed the modalities of facial expressions, acoustics, and transcribed utterances regarding their potential for emotion recognition during in-car speech interactions. We applied off-the-shelf emotion recognition tools for facial expressions and acoustics. For transcribed text, we developed a neural network-based classifier with transfer learning exploiting existing annotated corpora. We find that analyzing transcribed utterances is most promising for classification of the three emotional states of joy, annoyance and insecurity.

Our results for facial expressions indicate that there is potential for the classification of joy, however, the states of annoyance and insecurity are not well recognized. Future work needs to investigate more sophisticated approaches to map frame predictions to sequence predictions. Furthermore, movements of the mouth region during speech inter-

actions might negatively influence the classification from facial expressions. Therefore, the question remains how facial expressions can best contribute to multimodal detection in speech interactions.

Regarding the classification from the acoustic signal, the application of off-the-shelf classifiers without further adjustments seems to be challenging. We find a strong bias towards negative emotional states for our experimental setting. For instance, the personalization of the recognition algorithm (*e. g.*, mean and standard deviation normalization) could help to adapt the classification for specific speakers and thus to reduce this bias. Further, the acoustic environment in the vehicle interior has special properties and the recognition software might need further adaptations.

Our transfer learning-based text classifier shows considerably better results. This is a substantial result in its own, as only one previous method for transfer learning in emotion recognition has been proposed, in which a sentiment/emotion specific source for labels in pre-training has been used, to the best of our knowledge (Felbo et al., 2017). Other applications of transfer learning from general language models include (Rozental et al., 2018; Chronopoulou et al., 2018, *i.a.*). Our approach is substantially different, not being trained on a huge amount of noisy data, but on smaller out-of-domain sets of higher quality. This result suggests that emotion classification systems which work across domains can be developed with reasonable effort.

For a productive application of emotion detection in the context of speech events we conclude that a deployed system might perform best with a speech-to-text module followed by an analysis of the text. Further, in this work, we did not explore an ensemble model or the interaction of different modalities. Thus, future work should investigate the fusion of multiple modalities in a single classifier.

Acknowledgment

We thank Laura-Ana-Maria Bostan for discussions and data set preparations. This research has partially been funded by the German Research Council (DFG), project SEAT (KL 2869/1-1).

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In Václav Matoušek and Pavel Mautner, editors, *Text, Speech and Dialogue*, pages 196–205, Berlin, Heidelberg. Springer Berlin Heidelberg.

Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. 2015. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2).

Vladimir Andryushechkin, Ian Wood, and James O' Neill. 2017. NUIG at EmoInt-2017: BiLSTM and SVR ensemble to detect emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 175–179, Copenhagen, Denmark, September. Association for Computational Linguistics.

Hynek Boril, Omid Sadjadi, and John H. L. Hansen Hansen. 2011. UTDrive: emotion and cognitive load classification for in-vehicle scenarios. In *th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems, DSP 2011*.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Scott Brave, Clifford Nass, and Kevin Hutchinson. 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International journal of human-computer studies*, 62(2).

Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, pages 205–211, New York, NY, USA. ACM.

Alexandra Chronopoulou, Aikaterini Margatina, Christos Baziotis, and Alexandros Potamianos. 2018. NTUA-SLP at IEST 2018: Ensemble of neural transfer methods for implicit emotion classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 57–64, Brussels, Belgium, October. Association for Computational Linguistics.

Andrew Cooper and Konstantinos Vassilis Petrides. 2010. A psychometric analysis of the trait emotional intelligence questionnaire–short form (TEIQue–SF) using item response theory. *Journal of personality assessment*, 92(5):449–457.

Amy Coplan and Peter Goldie. 2011. *Empathy: Philosophical and psychological perspectives*. Oxford University Press.

Monique Dittrich and Sebastian Zepf. 2019. Exploring the validity of methods to track emotions behind the wheel. In Harri Oinas-Kukkonen, Khin Than Win, Evangelos Karapanos, Pasi Karppinen, and Eleni Kyza, editors, *Persuasive Technology: Development of Persuasive and Behavior Change Support Systems*, pages 115–127, Cham. Springer International Publishing.

Paul Ekman and Wallace V. Friesen. 1978. Facial action coding system: Investigator's guide. *Consulting Psychologists Press*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6.

Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3).

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark, September. Association for Computational Linguistics.

Figure Eight. 2016. Sentiment analysis: Emotion in text. Online. https://www.figure-eight.com/data/sentiment-analysis-emotion-text/.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Helen Harris and Clifford Nass. 2011. Emotion regulation for frustrating driving contexts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 749–752.

Klas Ihme, Christina Dömeland, Maria Freese, and Meike Jipp. 2018. Frustration in the Face of the Driver: A Simulator Study on Facial Muscle Activity during Frustrated Driving. *Interaction Studies*, 19.

Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Joint fine-tuning in deep neural networks for facial expression recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2983–2991.

Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee. 2016. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 10(2).

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 17–26, Vancouver, Canada, August. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Jonathan Klein, Youngme Moon, and Rosalind W. Picard. 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with computers*, 14(2).

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium, October. Association for Computational Linguistics.

Jinkyu Lee and Ivan Tashev. 2015. High-level feature representation using recurrent neural network for speech emotion recognition. In *Interspeech*. ISCA – International Speech Communication Association.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

James J. Lien, Takeo Kanade, Jeffrey F. Cohn, and Ching-Chung Li. 1998. Automated facial expression recognition based on facs action units. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 390–395.

Zhiyi Ma, Marwa Mahmoud, Peter Robinson, Eduardo Dias, and Lee Skrypchuk. 2017. Automatic detection of a driver's complex mental states. In Osvaldo Gervasi, Beniamino Murgante, Sanjay Misra, Giuseppe Borruso, Carmelo M. Torre, Ana Maria A.C. Rocha, David Taniar, Bernady O. Apduhan, Elena Stankova, and Alfredo Cuzzocrea, editors, *Computational Science and Its Applications – ICCSA 2017*, pages 678–691, Cham. Springer International Publishing.

Judi Mesman, Harriet Oster, and Linda Camras. 2012. Parental sensitivity to infant distress: what do discrete negative emotions have to do with it? *Attachment & Human Development*, 14(4).

Saif Mohammad and Felipe Bravo-Marquez. 2017a. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada, August. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017b. Emotion intensities in tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada, August. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2012. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3).

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3).

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June. Association for Computational Linguistics.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Clifford Nass, Ing-Marie Jonsson, Helen Harris, Ben Reaves, Jack Endo, Scott Brave, and Leila Takayama. 2005. Improving automotive safety by pairing driver emotion and car voice emotion. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, pages 1973–1976.

Michael Neumann and Ngoc Thang Vu. 2017. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. In *Interspeech*. ISCA – International Speech Communication Association.

Dimitri Palaz, Mathew Magimai-Doss, and Ronan Collobert. 2015. Analysis of cnn-based speech recognition system using raw speech as input. In *Interspeech*. ISCA – International Speech Communication Association.

Maja Pantic, Nicu Sebe, Jeffrey F. Cohn, and Thomas Huang. 2005. Affective multimodal human-computer interaction. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 669–676.

Timo Partala and Veikko Surakka. 2004. The effects of affective interventions in human–computer interaction. *Interacting with computers*, 16(2).

Maurizio Paschero, G. Del Vescovo, L. Benucci, Antonello Rizzi, Marco Santello, Gianluca Fabbri, and F. M. Frattale Mascioli. 2012. A real time classifier for emotion and stress recognition in a vehicle driver. In *2012 IEEE International Symposium on Industrial Electronics*, pages 1690–1695.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1.

Chris Pool and Malvina Nissim. 2016. Distant supervision for emotion detection using Facebook reactions. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 30–39, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Jonathan Posner, James A. Russell, and Bradley S. Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California, June. Association for Computational Linguistics.

K. Sreenivasa Rao, Shashidhar G. Koolagudi, and Ramu Reddy Vempada. 2013. Emotion recognition from speech using global and local prosodic features. *International journal of speech technology*, 16(2).

Alon Rozental, Daniel Fleischer, and Zohar Kelrich. 2018. Amobee at IEST 2018: Transfer learning from language models. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 43–49, Brussels, Belgium, October. Association for Computational Linguistics.

Klaus R Scherer and Harald G. Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2).

Klaus R. Scherer. 1997. Profiles of emotion-antecedent appraisal: Testing theoretical predictions across cultures. *Cognition & Emotion*, 11(2).

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark, September. Association for Computational Linguistics.

Björn W. Schuller, Manfred K. Lang, and Gerhard Rigoll. 2006. Recognition of Spontaneous Emotions by Speech within Automotive Environment. In *Tagungsband Fortschritte der Akustik – DAGA 2006*, pages 57–58.

Björn W. Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11).

Nicu Sebe, Ira Cohen, and Thomas S. Huang, 2005. *Handbook of Pattern Recognition and Computer Vision*, chapter Multimodal Emotion Recognition. World Scientific.

Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2012. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2).

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June. Association for Computational Linguistics.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Sujono and Alexander A. S. Gunawan. 2015. Face expression detection on kinect using active appearance model and fuzzy logic. In Widodo Budiharto, editor, *Proceedings of the International Conference on Computer Science and Computational Intelligence (ICCSCI 2015)*, volume 59 of *Procedia Computer Science*, pages 268–274. Elsevier.

Tessa-Karina Tews, Michael Oehl, Felix W. Siebert, Rainer Höger, and Helmut Faasch. 2011. Emotional human-machine interaction: Cues from facial expressions. In Michael J. Smith and Gavriel Salvendy, editors, *Human Interface and the Management of Information. Interacting with Information*, pages 641–650. Springer Berlin Heidelberg.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204.

Dimitrios Ververidis, Constantine Kotropoulos, and Ioannis Pitas. 2004. Automatic emotional speech classification. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages I–593–I–596.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592.