Literature Review and Summaries

Lecture on Emotion Analysis at University of Stuttgart
WS 2021/2022
https://www.emotionanalysis.de/
Roman Klinger

Authors: Linda Kim Greschner, Hannah Gabriella Hiergeist, Jan Hofmann, Hana Kang, Chin Liu, Sven Naber, Britta Schulz, Sakshi Shukla, Katharina Suhr, Nataliia Tkachenko, Itisha Yadav

Contents

Dialogue

Hannah Gabriella Hiergeist

> Sashank Santhanam, Samira Shaikh (2019): Emotional Neural Language Generation in Situational Contexts, Proceedings of the 4th Workshop on Computational Creativity in Language Generation. `https://aclanthology.org/2019.ccnlg-1.3`

Motivation

The motivation is to develop a language modeling approach that can generate an emotional content in dialogues, which take place in a given context. By applying transfer learning approaches to pretrained language models, appropriate emotional responses could be generated.

Data

The used data is the Empathetic-Dialogues corpus (Rashkin et al. 2019), which consists of dialogues in emotional situations. The dataset contains 32 emotion labels. The speaker chooses an emotional situation, to which the listener responds in an emotional appropriate manner. The trained corpus comprises of around 20,000 conversations and is tested with around 2,500 conversations.

Method

The researchers used the GPT-2 pretrained language model, which was then fine-tuned (model 1) and conditioned on a given emotion (model 2). The implementation was carried out by a PyTorch Transformer. For evaluating the quality of the responses, the researchers used BLEU and Perplexity (PPL) metrics. They also reported the length of the automatic responses, as well as the diversity in the responses. Two additonal reported metrics, which are important for the research, are the readability (linguistic quality of text) and the coherence (produced response is consistent with the topic). All these scores were compared to both baseline models proposed by Rashkin et al. (2019), the fine-tuned model and the model conditioned on emotions.

Main Result

Both models implemented by the researchers achieve a lower perplexity and a higher BLEU score than both baseline models proposed by Rashkin et al. (2019). Model 2, which was conditioned on emotions, produces more diverse responses and achieves a higher readability score than model 1, which was not conditioned on emotions. In a second evaluation step, 15 participants analyzed 25 random output sentences on three metrics: readabilty, coherence and emotional appropriateness. In all these categories, the participants rated the fine-tuned model (model 1) higher than the emotion-conditioned model (model 2).

Critical Reflection, Limitations

In my opinion, the implementation of a language modeling approach that can generate emotional content is very important and therefore extremely interesting. The evaluation of both models in contrast surprises me though, because the automated metrics (readabilty, coherence and diversity) are higher on the emotion-conditioned model, whereas the human ratings of the same parameters are higher on the fine-tuned model. This implies, that the emotion-conditioned model is perceived less accurate by humans in terms of the quality of the responses. It would also be very important to know the ratings of emotional appropriateness in general, because that would be an interesting factor to evaluate the models. The results also show that the emotion-conditioned model achieves just minimal higher ratings than the fine-tuned one and also slightly worse ratings in the human evaluation. The researchers could adress this in a more specific evaluation of the models.

Hana Lin Kang

> Deeksha Varshney, Asif Ekbal, Pushpak Bhattacharyya (2021): Modelling Context Emotions using Multi-task Learning for Emotion Controlled Dialog Generation, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2919–2931.
>
> `https://aclanthology.org/2021.eacl-main.255.pdf`

Motivation

This work aims to produce emotion controlled responses for multi-turn conversations using relevant context knowledge and emotion labels. Its motivation is to develop an emotion-controlled dialog generation model, which generates responses according to the target emotion style, while trying to balance language fluency and emotion quality in the generated responses.

Data

This research evaluates the proposed model on the Topical Chat dataset from Gopalakrishnan et al.(2019), which is based on a knowledge-grounded human-human conversation where the underlying knowledge spans 8 broad topics. This dataset consists of two types of files: the conversation files contain a conversation between two workers on Amazon Mechanical Turk, and the reading-set files contain knowledge sections from different data sources (Wikipedia, Washington Post...) that are served up to a particular Turker to read and refer to as they are having a conversation.

Method

Self-attention (Vaswaniet al., 2017) based encoder: for computing the context features in a dialog. GRU network: for addressing the previous context of utterances in a multi-turn conversation. Multi-task learning: for encoder to learn common and prominent features in the input sequence. Focal Loss (Lin et al., 2017): for addressing imbalance between the emotion classes during training. Consistency Loss: for reducing the difference between the attention weights from different tasks.

Main Result

This paper proposed a new deep learning framework for modeling emotion-grounded conversations using emotion labels as the guiding attributes. The proposed model with appropriate loss functions can ensure emotional relevancy of the generated response and therefore improves user engagement. The automatic evaluation shows that this model improves on the F1 and BLEU metrics significantly, while the human evaluation results reveal that this method improves not only the fluency and adequacy scores but also the emotional accuracy scores.

Critical Reflection, Limitations

In this research, the idea of using both the automatic and human evaluation methods is great. Because in this way, not only the results of the most well-known metrics, like BLEU, F1, perplexity and n-gram diversity can be considered, but also measuring of the quality of the generated text from a human perspective could be taken into account. However from the perspective of reducing errors in research, there is still room for improvement in this work, e.g. the issues of repetition, common phrases and emotional inconsistencies. In addition, although the grammatical correctness of the proposed model is almost perfect and its contextual relevancy also quite good, but the emotional accuracy with the scores 0.80/0.60 indicates that the predicted responses do not always represent the correct emotion. Last, using pre-trained language models might be the next step for this work.

Sakshi Shukla

> Wang, Zhang et al. (2020): Contextualized Emotion Recognition in Conversation as Sequence Tagging, SIGDIAL. `https://aclanthology.org/2020.sigdial-1.23`

Motivation

CESTa models Emotion Recognition in Conversation task as a Sequence Tagging task to preserve the emotional consistency in a conversation.

Data

The CESTa model is performed on three different datasets. IEMOCAP and MELD are multimodal data whereas DailyDialogues is textual data.

IEMOCAP (Interactive Emotional Dyadic Motion Capture) with emotions: happy, sad, neutral, angry, excited and frustrated. DailyDialogue, daily life based data with emotions: anger, disgust, fear, happiness, sadness, surprise and other. MELD (Multimodal Emotion Lines Dataset) is collected from TV series Friends including happy/joy, anger, fear, disgust, sadness, surprise and neutral as labels.

Method

ERC predicts emotion tags locally, to enhance it to Sequence Tagging, each tag predicted should have a relation with the neighbouring tags. The best global tag out of the list of predicted tags is selected. The architecture of the CESTa model is divided into four major categories: Utterance Feature Extraction, Global Context Encoder, Individual Encoder and Conditional Random Field(CRF) Layer.

The textual feature is extracted by a single-layer CNN in the Utterance Feature Extraction and fed into the global and individual context encoders which learn inter speaker and self dependency in Global Context Encoder and Individual Context Encoder. The concatenation of the global context encoding and individual context encoding is fed into the CRF layer. While decoding, the tag sequence with maximum score is considered.

Main Result

CESTa outperforms the-state-of-art model DialogueGCN by 3% on IEMOCAP dataset. For MELD dataset, CESTa achieves better results than the baselines. The evaluation metric used for IEMOCAP and MELD is weighted macro F1. For DailyDialogues dataset, the model accounts with 63.12 micro-averaged F1 and outperforms baseline models with a large margin. The results show that ERC preserve context when used as a Sequence Tagging task.

Critical Reflection, Limitations

The paper is written in a comprehensible way and clearly explains all the four methodologies used to built CESTa. The explanation to use CRF for preserving the context is well defined. Each and every model used in CESTa is properly designated with appropriate reasoning and purpose of application. This particularly helps in analysing the complete processing of the CESTa model.

The reasons why CESTa was incompatible to some cases with all the three datasets is well drafted. This also leads to scope of improvement for future possibilities.

The limitation of this model is that it is dependent on neighbouring emotion tags to yield the global emotion tag. Hence, this system is not valid on online dialog system which does not have any future emotion.

Itisha Yadav

Peixiang Zhong, Di Wang, Chunyan Miao (2019): Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. `http://aclanthology.lst.uni-saarland.de/D19-1016.pdf`

Motivation

The paper proposes the use of transformers with an external knowledge base over RNN, CNN and baseline bert. The self-attention and cross-attention proposed in the paper captures intra-sentence and inter-sentence correlations. Unlike base BERT model which only uses an encoder, it separates context and response into encoder and decoder architecture. To get a better semantic representation of each word they derive a context-aware, emotion related common sense knowledge from ConceptNet and NRC-VAD lexicon. Furthermore, for efficient exploitation of context, a hierarchical self-attention approach is used.

Data

The model is trained and evaluated on five emotion detection data-sets of different sizes.

Dataset	Description	Labels
EC	Tweets	happiness, sadness, anger and other
DailyDialog	Human written daily communications	neutral + Ekman's six basic emotions
MELD	Scripts from TV show *Friends*	neutral + Ekman's six basic emotions
EmoryNLP	Scripts from TV show *Friends*	peaceful, sad, powerful and other
IEMOCAP	Emotional Dialogues	happiness, anger, frustrated and other

Method

The system is broadly divided into two parts, encoder handles the context and the decoder handles the response. Each conversation consists of utterances and each utterance is made up of tokens. Similar concepts of the token are derived from the Knowledge Base (ConceptNet). Utterances consisting of word embedding and concept embedding are filtered using Dynamic Context-Aware Affective Graph Attention mechanism which computes the relatedness score with the complete context of the conversation and affectiveness score, which is the emotion intensity score. Concepts not included in NRC-VAD data-set are discarded. Hierarchical approach is used for better modelling of the context by applying self-attention at two levels, first at utterance level and then at the context level. The final context and response representations are combined by applying cross-attention technique. The combined representation then goes through a max pooling layer and a softmax function to get the final result.

Main Result

The paper concludes that including external knowledge base and NRC-VAD lexicon improved the overall results when compared with the baseline models like cLSTM (contextual LSTM), CNN, CNN+cLSTM, BERT-BASE and DialogueRNN(state-of-the-art). The model performed well on all datasets except IEMOCOP, adding more context is leading to diminishing performance gain.

Critical Reflection, Limitations

Though the model has shown improved overall performance, but it fails on emotions like fear and disgust, maybe because internally, in terms of VAD both the emotions are same. The scarcity of data for these emotions can also be a problem which needs to be handled.

Face

Sven Naber

> Mostafa Shahabinejad, Yang Wang, Yuanhao Yu, Jin Tang, Jiani Li...(2021): Toward Personalized Emotion Recognition: A Face Recognition Based Attention Method for Facial Emotion Recognition, 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). `https://ieeexplore.ieee.org/document/9666982`

Motivation

A difficulty in facial emotion recognition is that the facial expression of emotions has considerable individual differences. As the space of possible facial expression is determined by the physical properties of an individual the realisation of distinct emotions too will be influenced by these factors. In facial recognition (FR) tasks there is a strong focus on these subtle individual differences, additionally state-of-the-art FR systems already boast a high accuracy (98%+). FER could benefit from the information of a FR model.

Data

The CNNs of both the FR and FER task are based on the EfficientNetB0 model, which was trained on ImageNet. As training data they used the training sets of AFEW (773 images) and the categorical model of AffectNet (287k images), which have the 8 categories anger, disgust, fear, happy, neutral, sad, surprise, and contempt (not in AFEW and the 7 categories variant of AffectNet). As test data they used the validation sets (AFEW 383 images and AffectNet 8k).

Method

Both the FR-CNN and FER-CNN use EfficientNetB0 as basis. The FR-CNN is frozen while the FER-CNN is fine-tuned with the training data. A spatial attention mask is generated by applying first a 2D convolution, then the sigmoid function and lastly a 2D resizing operation on the feature map (of a low level n) of the frozen FR-CNN. This attention mask is then applied via a spatial channel-wise multiplication to the feature map F (at level m) of the FER-CNN (resulting in a new feature map H). An element-wise maximum operation on F and H then generates the ”individualized” feature map which the FER-CNN further uses. The resulting models were evaluated on the test data.

Main Result

The proposed FR attention mechanism improves the performance of the FER model. Improvement of accuracy by 0.69% (AffectNet 7 classes, to 66.4%), 1.56% (AFEW, to 60.83%) and 1.96% (AffectNet 8 classes, to 63.28%) compared to prior state-of-the-art.

Critical Reflection, Limitations

FER with prior unseen faces works under the assumption that there are commonalities in the emotion expression of individuals. This implies that there must be commonalities in the information the brain sends to the muscles to create different facial expressions. How this information is translated in facial expressions is dependent on individual physiology - mainly the shape and structure as well as the muscle distribution of an individuals face. This also creates additional subject variation. FR capitalizes on these individual differences. But the information used for FR may encompass some but not all the information needed to explain physiological variation. Maybe a model that tries to predict for a given face either a relaxed facial expression or one which defines the space of physiologically possible facial expressions could provide better information for FER.

Britta Schulz

> Ng, Nguyen, et al. (2015): Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning, ICMI '15. `https://dl.acm.org/doi/abs/10.1145/2818346.2830593`

Motivation

The paper describes and analyses the submission the authors made to the 2015 "EmotiW" contest, for the sub-challenge "Static Facial Expression Recognition in the Wild". Investigating improvements of deep Convolutional Neural Network (CNN) architectures and training methods for facial expression analysis is very important since there are many important applications in different research area intersections. For instance, it is applied in HCI, surveillance or crowd analyses. The given dataset holds close-to-real-life conditions as the faces are extracted from movies where illumination conditions vary and occlusions or motion blur occur. The challenge's task has been to classify the expressions neutral, happy, surprised, fear, angry, sad, and disgusted.

Data

Initially, they start with a pretrained CNN having used ImageNet data. On that starting point, they applied two fine-tuning steps; firstly using the FER-2013 datset, which is very large, but with a much lower resolution of only 48x48 instead of 256x256 pixels. Subsequently, they applied a second fine-tuning step using EmotiW's training data, before evaluating with EmotiW's development and test data.

Method

As already initiated in the previous section, the autors used a transfer learning approach for deep CNNs and a two-staged fine tuning using the large but low-resolutional auxiliary data provided by FER-2013. Regarding data preparation, they used a different approach than the baseline paper by getting a squared face bounding box that focuses on the forehead area rather than the chin, based on previous research results.

Main Result

Evaluating the method has shown that it yields a significant accuracy increase of 16% compared to the baseline. A very interesting and surprising result is that a variation that only used FER-2013 data for training still achieved an improve of close to 15% compared to the baseline. Since the images in that dataset have not been aligned and a much lower resolution, it follows that quantity of data is probably much more important than quality for training such a CNN. The irrelevance of resolution is also an interesting parallel to the neuroscientific finding that the human brain processes faces holistically.

Critical Reflection, Limitations

Even though the paper only achieved the challenge's third place, it has been by far the most influential. The reason might be that they do not show complicated architectures but the significant effect of using additional, much larger training data, which means considerably less effort. Nevertheless, the achieved accuracy of 55.6% on the test set is still not stunning. Some ambiguous or subtle expressions are still very hard to classify, especially when they only occur on a small percentage of the training samples. Accordingly, it is very important to extend this paper's method with better architectures as well, for example.

Nataliia Tkachenko

> Heaven D., (2020): Why faces don't always tell the truth about feelings, Nature. `https://media.nature.com/original/magazine-assets/d41586-020-00507-5/d41586-020-00507-5.pdf`

Motivation

Is Ekman's suggestion, that around the world, humans could reliably infer emotional states from expressions on faces, really true?

Data

The author takes an overview on research literature from classical studies (Ekman) to the newest ones.

Method

The author depicts a wide literature review above resent studies.

Main Result

Researchers are increasingly split over the validity of Ekman's conclusions. There are studies that show that there is little to no evidence that people can reliably infer someone else's emotional state from a set of facial movements. There is also variation in how faces express emotion. The researchers need to observe more, like Darwin did for *On the Origin of Species*. They should watch what people actually do with their faces and their bodies in real life and not just observe emotions in the lab. And then use machines to record and analyze facial expressions.

Critical Reflection, Limitations

The article is from 2020 and shows recent research findings. Unlike other articles, it questions the reliability of classical studies about emotions. A lot of models are often based on Ekman's six emotions (Happiness, Sadness, Anger, Fear, Disgust and Surprise) which count as universal ones. Ekman's findings contribute to the understanding of facial expressions and their corresponding emotional states. His research had a huge influence on the development of the emotion detection in faces. Some algorithms are broadly used in marketing agencies, designing social robots or smart environments in hospitals. But some algorithms are also being used for job interviews and at borders or legal settings, where inexactness can have a strong impact on someone's personal life. So it is important to discuss if algorithms are reliable and if they are based on right conclusions. The author of the article writes that facial expressions vary widely between contexts and cultures. I could also learn from the article that there are some differences in how Westerners and East Asians show their emotions through facial expressions. For me, the article mentions aspects that make me doubt if faces are a reliable indicator of one's feelings. Different people express their emotions differently according to their personal traits. It is also obviously possible to experience emotions without showing them on a face. People can also fake emotions and experience feelings without moving their faces. According to this, the automatic facial expression detection might be easily tricked. The article is a little bit limited in regard to examples of positive outcomes from some emotion detection in faces. It could be improved by giving some more examples of successful models and show some more numbers according to it. The article opens up other perspective on to the topic. I think that reflecting and questioning the concept of emotion detection is an important step in the further development of it.

Mental Health

Lynn Greschner

> Kyo-Joong Oh, DongKun Lee, ByungSoo Ko, Ho-Jin Choi (2017): A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation, IEEE 18th International Conference on Mobile Data Management. `https://ieeexplore.ieee.org/abstract/document/7962482`

Motivation

While the need for psychiatric therapy and/or counseling for people with mental health issues, such as anxiety, depression, or lethargy, increases, the availability of such therapies (by experts) is low and expensive. As talking or interacting with someone can highly improve the mental health of a person (especially in emergencies such as suicidal thoughts), the authors argue that providing a conversational service (chatbot) is needed. Implementing emotion analysis with multi-modal methods allows to grasp emotional states of users and leads to an improvement in counseling and eventually improves the mental health state of persons (including healthy changes in self-awareness, personality, and behavior). To construct the conversational service for psychiatric counseling, they concern three different aspects: A: Natural Language Understanding; B: Emotional Dialogue Analysis; C: Sentence Generation for Psychiatric Counseling.

Data

A. Collected from Korean Wikipedia, Namuwiki, and news articles. Contains 49,864,477 sentences. They trained 200 dimensional word vectors of more than 380,000 vocabularies.
B. Training data was collected from various media sources (dramas and radio), containing emotional information. It is in the from of voice, video, text and bio-sense.
C: Not mentioned.

Method

A: GRU for sentence similarity analysis. Analyze domains with high relevance to input sentences.
B: Emotional model which can express more than eight kinds of emotions for emotion recognition, learning, and inferencing (w.r.t. Emotion wheel by Plutchik).
C: Technique based on RNN-decoder to generate natural sentences w.r.t. the context qualities of the conversation. Point Network Model to learn important keywords in sentences.

Main Result

The authors do not report performance results. However, they argue that understanding and recognizing emotions of users improves psychological counseling with chatbots.

Critical Reflection, Limitations

Using a chatbot to assist people with mental health issues is a great approach. The motivation of this paper hints towards an interesting idea: using emotion analysis to allow chatbots not only to understand the current emotional state of the user but, additionally, to adapt the responses to be empathic or adapted to the emotional need of users seems very promising. However, the paper is difficult to read (typos, unclear sentences (maybe translation errors?)) and does neither explain the models well, nor report any statistical results or performance results of the models, and the authors also do not explain from which sources they obtained their data. Still, the paper can influence researchers to work on implementing emotion analysis to chatbots in order to provide a helpful interactive system to improve mental health issues.

Chin Liu

> Jianqiang Xu, Zhujiao Hu, Junzhong Zou, Anqi, Bi . . . (2019): Intelligent Emotion Detection Method Based on Deep Learning in Medical and Health Data, IEEE Access. `https://ieeexplore.ieee.org/document/8937486`

Motivation

People suffering from mental health issues are increasing nowadays, thus, it is important for people to be able to recognize and cope with negative emotions in time. However, it is not easy for everyone to recognize what kind of human fatigue he/she is suffering and take corresponding actions when feeling fatigued. Hence, the authors proposed three types of emotional fatigue: 1. physiological fatigue: fatigue caused by physiological exhaustion 2. repetitive fatigue: fatigue caused by continuous repetitive actions 3. environment fatigue: fatigue caused by continuously working in a closed space, and an emotion detection method based on deep learning in medical and health data.

Data

In the study, two data are used:

- Electrocardiograms (ECG) data features are collected from three sensors of wearable devices.
- Emotional text features from communication software.

These data are collected for ten days from 10 volunteers in total (5 males and 5 females) through mobile phones and wearable devices.

Method

A multi-channel convolutional aotoencoder neural network is proposed to design:

- ECG-MCAE network structure: learn the ECG characteristics of time series data
- Text-CAE network structure: learn the emotional text features

In order to let the network model learn from various feature representations, different feature information is combined. The reason why CNN is chosen is that it has advantages in processing time series data and learning data-driven features.

Main Result

The proposed model achieves an average accuracy of over 85% in predicting emotional fatigue. When emotional fatigue is detected, users are given personalized feedback (e.g. listen to favorite music, taking a rest, and go for a walk) according to the type of emotional fatigue.

Critical Reflection, Limitations

First of all, mental health problems are definitely one of the most important issues in today's society. Most people may not recognize themselves being in the state of a physical fatigue or psychological fatigue, thus the proposed method may be a reminder and prevention for all of us. However, there are some limitations. As the authors mentioned, the number of the collected sample fatigue states are different and it caused a great difference on the performance of the results, thus the process of sample selection can be improved. Further, the medical and health data definitely play an important role in detecting fatigue, however, I expect more data other than only ECG data. Considering other aspects of medical data would improve the accuracy and reliability of the proposed method.

Speech

Jan Hofmann

> Aftab, Morsali, et al. (2021): Light-SERNet: A lightweight fully convolutional neural network for speech emotion recognition, arXiv preprint arXiv:2110.03435. `https://arxiv.org/abs/2110.03435`

Motivation

Creating an efficient and lightweight model for speech emotion recognition while maintaining the performance of state-of-the-art models.

Data

The authors of this paper apply their model to the datasets IEMOCAP (Interactive emotional dyadic motion capture database, 5531 samples, English-language, *angry*, *exiting*, *happiness*, *sadness* and *natural*) and EMO-DB (Berlin Database of Emotional Speech, 535 utterances, German-language, *anger*, *boredom*, *disgust*, *fear*, *happiness*, *sadness* and *natural*).

Method

The model architecture proposed in this paper consists of three parts: The input pipeline block, feature extraction (Body) blocks and classification block (Head).

The input pipeline is calculating Mel frequency cepstral coefficients (MFCC) from the audio input. The feature extraction block is further divided into two parts: The first part consists of three parallel 2D-convolutions which are designed with specific kernel sizes to extract spectral, temporal and spectral-temporal dependencies when applied to the MFCC inputs. The second part of the feature extraction block uses the concatenated output of the first part and consists of multiple consecutive local feature learning blocks (LFLB). One LFLB consists of a convolutional layer, a batch normalization layer using ReLU activation and average pooling. The classification block uses the output of the last LFLB and includes dropout and a fully connected layer with softmax activation.

The authors of the paper apply their model to the above mentioned datasets and report experimental results using 10-fold cross-validation.

Main Result

The main result of this paper is the architecture of a convolutional DNN for speech emotion recognition which is smaller in size while reaching almost the same (IEMOCAP) or higher recognition performance (EMO-DB). On EMO-DB, for example, the authors report 94.14% macro F1.

Critical Reflection, Limitations

Overall, the paper is written in a clear and understandable way. Hyperparameters and the proposed model architecture are stated and described with just the right detail enabling a straight forward implementation and reproducibility. Sometimes, however, specific architectural decisions could be more specific. I'm, for example, curious why the authors changed the activation in their LFLB block, which is inspired by another work, from exponential linear unit (ELU) to rectified linear unit (ReLU) and max-pooling to average-pooling.

This work is a good reminder that one should be very thoughtful when designing artificial neural networks in order to make use of their full potential. While other state-of-the-art approaches in speech emotion recognition often use combinations of CNNs, RNNs (mostly LSTMs) and attention using much greater computational complexity it is very interesting to see such a lightweight CNN architecture outperform more sophisticated models and frameworks.

Katharina Suhr

> Michael Braun, Jonas Schubert, Bastian Pfleging and Florian Alt (2019): Improving Driver Emotions with Affective Strategies, Multimodal Technologies and Interaction 3.1. `https://www.mdpi.com/2414-4088/3/1/21`

Motivation

Driving while in a negative emotional state can lead to unsafe driving, which also leads to decreased safety for the driver and other participants on the road. Since a car is a small enclosed spaces the researchers wanted to find out which influencing strategies could change the drivers emotions in a positive way.

Data

In this study there was no data involved. The paper does start with a selection of influencing strategies, that were found in other papers. These strategies where narrowed down to four strategies by twelve UI/UX researchers of the BMW Group to get a starting point. These four strategies were: Ambient light, Visual Notification (graphic visualization of their current state), Voice Assistant, Empathetic Assistant

Method

To figure out which of the influencing strategies work best in influencing the emotional state of a user in a positive way, a user study was conducted. A total of 60 people (42 male, 18 female) did participate in the study. This user study took place in a BMW driving simulator. To create the emotions in a participant they needed to recount a story from their own life to evoke the desired emotion. Than the strategy was started and the difference was measured. To measure the driving performance the researcher collected for example the information about the car's position within the lane, eye tracking, the drivers workload and the strength of their emotion.

Main Result

The user study showed that sad drivers had imperfect performance, while angry drivers showed a significantly worse performance. The most effective influencing strategy is the empathetic assistant, followed by the voice assistant. One feedback that both assistants got was that they felt too patronizing. But still both the assistants were preferred to more subtle strategies like Ambient light. Compared to Visual Notification, which was not perceived very well, Ambient Light was still perceived quite well. So their conclusion of what works best is: A natural voice interaction with subliminal cues (e.g., light), that focuses on empathy, without patronizing the user.

Critical Reflection, Limitations

A lot of the limitations in my opinion come from the set up. The use of a simulator will alter the reactions of the participants. I am not sure how well the recounting of events works to recreate emotions. In addition they also mentioned that in real world the workload can additionally influence the emotional state of a person. So maybe in some cases the workload of a voice assistant can alter the state of a person in a negative way. For a first run 60 participants are probably sufficient, but it would definitely be interesting to see the results of a larger study. In this larger study I would also expect the distribution of gender to be more even then in this study. Additionally a small survey to the driving background (how often do they drive, how long do they have their license, etc.) of the participants could be informative.