Research Article

Thomas Schmidt*, Miriam Schlindwein, Katharina Lichtner, and Christian Wolff

Investigating the Relationship Between Emotion Recognition Software and Usability Metrics

https://doi.org/10.1515/icom-2020-0009

Abstract: Due to progress in affective computing, various forms of general purpose sentiment/emotion recognition software have become available. However, the application of such tools in usability engineering (UE) for measuring the emotional state of participants is rarely employed. We investigate if the application of sentiment/emotion recognition software is beneficial for gathering objective and intuitive data that can predict usability similar to traditional usability metrics. We present the results of a UE project examining this question for the three modalities text, speech and face. We perform a large scale usability test (N = 125) with a counterbalanced within-subject design with two websites of varying usability. We have identified a weak but significant correlation between text-based sentiment analysis on the text acquired via thinking aloud and SUS scores as well as a weak positive correlation between the proportion of neutrality in users' voice and SUS scores. However, for the majority of the output of emotion recognition software, we could not find any significant results. Emotion metrics could not be used to successfully differentiate between two websites of varying usability. Regression models, either unimodal or multimodal could not predict usability metrics. We discuss reasons for these results and how to continue research with more sophisticated methods.

Keywords: Affective computing, usability engineering, usability, sentiment analysis, emotion analysis, usability test, system usability scale

***Corresponding author: Thomas Schmidt,** Media Informatics Group, University of Regensburg, Regensburg, Germany, e-mail: thomas.schmidt@ur.de

Miriam Schlindwein, Katharina Lichtner, Christian Wolff, Media Informatics Group, University of Regensburg, Regensburg, Germany, e-mails: miriam.schlindwein@stud.uni-regensburg.de, katharina.lichtner@stud.uni-regensburg.de, christian.wolff@ur.de

1 Introduction

One of the most established definitions to describe the concept of the usability of a software is described in the DIN EN ISO 9241: the degree to which a user of the software can achieve their goals effectively, efficiently and with high levels of satisfaction [13]. One of the most popular methods to evaluate and optimize usability is the usability test [2] in which an adequate number of participants try to solve typical tasks of a software. Metrics like time, task completion rate or subjective assessments via questionnaires are used to interpret the usability of the product [2]. Another factor researchers as well as usability professionals are often interested in is the emotional state of the user [1, 7, 8, 46]. Researchers currently either use common questionnaire-based metrics from psychology [1, 7, 46] or develop their own, adapted to human-computer interaction (HCI) issues [8]. More objective and intuitive is the usage of physiological measurement instruments [46]. The majority of physiological measurement instruments are challenging to apply for simple usability tests (e. g. measuring heart rate, electrodermal activity) since they require a sophisticated technical set-up. However, advances in affective computing have led to the availability of easy-to-use general purpose emotion recognition tools for various modalities like text [23], speech and faces (cf. Daily et al. [10]). There is little usage of these tools in the context of usability engineering so far. Researchers explore the benefits of such tools in usability engineering [20] or investigate to what extent the output of these tools correlates with more common usability and user experience metrics [48] – with ambivalent results. We investigate if using these tools is beneficial for usability engineering since they offer an objective and intuitive metric that cannot be skewed by the users' subjective opinions and memory bias as with common questionnaire based methods. Furthermore, these tools can be applied automatically and might allow to save up time that is used on lengthy questionnaires and other methods. To investigate our proposition, we have performed a large-scale comparative usability test (N = 125) with two websites of varying usability and three sentiment/emotion recognition approaches based on three different modalities. We

analyze (1) if the output of these approaches correlates with more traditional usability metrics, (2) if we can successfully differentiate software of varying usability solely with these metrics and (3) if the output can successfully predict usability metrics.

2 Related Work

There is a plethora of similar as well as diverging definitions of emotions and sentiment concepts in psychology [22]. Mulligan and Scherer [27] propose a working consent definition for the term emotion that can be summarized as an emotion being an affective intentional episode triggered and guided by appraisal that also consists of bodily changes that are felt. The circumplex or valence-arousal model by Russell [35] defines emotion on a two dimensional scale consisting of valence and arousal, and different emotions can be situated on different parts of these scales. Other concepts define separated main categories, e. g. in Plutchik's [31] famous wheel of emotions of the eight basic emotions (joy, trust, fear, surprise, sadness, disgust, anger, anticipation). These emotion categories can also be assigned with various intensity levels. Agarwal and Meyer [1] argue that sadness, anger, surprise, fear, happiness and disgust are the six commonly accepted classifications of emotions. However, Schröder et al. [41] offer up to 48 separate emotion classes when defining an XML standard for emotion representation. The term sentiment, while often used interchangeable to affect and emotion, is used to describe if and to what extent a person likes or dislikes something or how something causes positive or negative emotions [5].

The role of emotions and sentiment have gained a lot of interest in HCI in recent years [5, 10, 45]. The research area concerned with the role of human emotions in HCI is nowadays often referred to as affective computing [45]. Two of the current major areas of this research branch are (1) the simulation of emotions in machines e. g. in robot-human interaction (e. g. Suzuki et al. [43]) and the computational analysis, detection and reaction towards human emotions and sentiments in various systems (cf. Tao & Tan [45]; Daily et al. [10]). With the advent of powerful machine learning technologies, this research area has shown significant progress in the last decade [10]. The analysis and detection of human emotions rely on various forms of human cues, most importantly facial and speech-based cues [19]. Other cues that have been examined are physiological data (e. g. galvanic resistance) or body posture [19]. Of the possible modalities, facial emotion recognition is considered as being the most robust technique, achieving accuracies above 90 % for standardized corpora of human faces [32]. Speech analysis, on the other hand can achieve similar accuracies but is on average regarded as inferior (cf. Hudlick [19]; Saxena et al. [38]). The popularity of these methods has led to multiple commercial and non-commercial tools like *Affectiva* [25] or *OpenFace* [3]. Examples for speech emotion recognition are *Vokaturi*[1] and *BeyondVerbal*[2] (cf. Garcia-Garcia et al. [16]).

To describe the analysis of text with respect to sentiment and emotions, the terms opinion mining, sentiment and emotion analysis are frequently used [23, p. 1]. Textual sentiment analysis is mostly focused on application areas like social media and online reviews (cf. Mäntylä et al. [24]) but is also gaining popularity in computational literary studies (cf. Schmidt & Burghardt [39]; Kim & Klinger [21]). Furthermore, in contrary to facial and speech emotion recognition, textual sentiment analysis focuses on the analysis and detection of valence (also often referred to as polarity), meaning whether the opinion or sentiment expressed in a text towards something is positive, negative or neutral. Concerning the methods used, various forms of machine learning approaches are applied. However, for more general purpose tasks or if training corpora for machine learning are missing, there are also rule-based approaches that can be applied (mostly referred to as dictionary- or lexicon-based sentiment analysis; Taboada et al. [44]). Textual sentiment analysis can achieve above 90 % accuracy on text sorts like product reviews [47].

Next to these computational approaches, research in psychology offers a multitude of possibilities for measuring emotions that can be structured into verbal, nonverbal and physiological measurement instruments [30]. Physiological measurements are to some extent in line with the major computational approaches e. g. when specific facial expressions are measured. There are, however, other common methods like measuring heart rate, electrodermal activity or pupil responses [30]. Most common in usability engineering are various verbal and nonverbal instruments to measure emotions [1, 8, 18, 46]. Verbal instruments are typically short questionnaires in which people indicate their emotions and feelings e. g. via Likert-scales or semantic differentials (cf. Agarwal and Meyer [1]; Bruun et al. [7]). Nonverbal instruments use visual representations to measure emotional state, e. g. smiling faces. The motivation for nonverbal measurements is to avoid problems because of language or cultural differences. Two

1 https://vokaturi.com/

2 https://beyondverbal.com/

of the most popular nonverbal measurement tools are the *Self-Assessment Manikin* (SAM; Bradley & Lang [4]) and *EmoCard* [11]. These verbal and nonverbal measurements are usually the method of choice in usability engineering to operationalize emotion and sentiment. However, for a long time the majority of established usability questionnaires has focused on the key factors efficiency, effectiveness and satisfaction only (e. g. Brooke [6]), while questionnaires in the context of user experience tend to integrate items to measure and rate emotional states or attribution e. g. *UEQ* [40] or *AttrakDiff* [17]. Champney and Stanney [8] try to eliminate the lack of emotion measurement instruments for usability evaluation by developing the *emotional profiling method*, which connects the elicitation of emotions with specific attributes of a product. Thüring and Mahlke [46] perform a usability test to investigate which type of emotional responses are influenced by usability. They evaluate this question not only via questionnaire items like the SAM, but also with physiological reactions connected to emotional. While *SAM* showed significant differences connected to usability, the only physiological factor showing significant and consistent effects was the heart rate. Agarwal and Meyer [1] argue that emotional response is an integral part of usability and user experience. They compare the results of usability metrics and of measurement tools in a traditional usability study. The results show that the two interfaces they compare have no significant differences considering usability metrics (time on task, number of errors), however the measurements applied to gather emotional attribution, here via *EmoCards* [11] and the *PAD* semantic differential [26], did show significant differences. Hassenzahl and Ullrich [18] highlight some problems of acquiring emotions via items in questionnaire that are predominantly given out at the end of the test. They developed an approach to gather more immediate emotion ratings after smaller steps of interaction tasks. Bruun et al. [7] argue that there is still a relevant memory bias for verbal emotion responses. In their study, they could show significant correlations between emotional states measured via *galvanic skin response* (GSR) and *SAM*-ratings given out via cued-recall instead of after every task. Cued-recall means here that participants first watched video snippets of their usability tests before making the *SAM* rating.

Despite positive results, verbal and non-verbal measurement tools via questionnaires are prone to subjectivity and extend the experiment time when applied. On the other hand, the application of advanced methods to measure physiological attributes are challenging to implement since they often require an elaborate set of hardware. Easy-to-use emotion recognition software is hardly used in practical usability engineering today. Recently, however, Xu et al. [48] have used, among other methods, a face recognition software to compare two different in-vehicle information systems in the context of automotive usability engineering. They have used a variety of measurement methods and among others, valence and arousal values were measured via facial emotion analysis. While they showed several significant differences via traditional methods, concerning the emotion metrics, only arousal showed a significant difference for various tasks. These differences were also in line with subjective satisfaction ratings while the valence counterintuitively did not agree with the subjective ratings. Johanssen et al. [20] argue that emotion recognition software might help not just in predicting the emotional state and overall usability but concrete usability issues. They have developed a framework for interactive mobile applications to use facial emotion recognition during usability tests: *EmotionKit*. Via qualitative analysis they have identified that the results of the plots of emotion progressions match their manual observations and that combined with the tracking of specific interface events one can identify usability problems more easily (e. g. by looking at strong deflections of negative emotions at specific events). By summing up all emotion channels, they show that this overall emotion response is mostly in line with notes made via traditional manual observation. Summing up the research in this area, results about the application of emotion recognition software are ambivalent considering the relationship with usability metrics [46, 48].

3 Research Questions

We are primarily interested in analyzing if and how emotion recognition tools and methods of various modalities can be used to automatically predict the subjective and objective usability in usability testing. We will first investigate if the output of emotion recognition correlates with usability metrics. As usability metrics we select the *task completion rate* and the *System Usability Scale* score (SUS; [6]):

RQ1: Is there a relationship between the output of sentiment/emotion recognition tools and usability metrics?

Furthermore, we will analyze if sentiment/emotion recognition tools can be used to differentiate tools of varying usability in usability tests (similar to the way usability metrics are used).

RQ2: Is there a difference between the outputs of sentiment/emotion recognition tools when comparing two products of varying usability?

Finally, we will perform regression analysis with the output of sentiment/emotion recognition tools to investigate if models based on these factors can successfully predict the outcome of usability metrics.

RQ3: Can regression models based on the output of sentiment/emotion recognition tools predict usability metrics?

4 Methods

We have conducted a usability test for two E-shop websites of the same type. The overall test is similar to the well-established method of *Guerilla Usability Testing* [28]. The usability test is structured by tasks and we applied the "thinking aloud"-method. The study follows a within-subject design and is counterbalanced to avoid learning effects. We gathered several established usability metrics and recorded the tests: (1) the desktop interaction and (2) the video of the webcam and a microphone.

4.1 Test Object

To validate our research question we argue that it is necessary to induce variance concerning usability metrics which is more likely by evaluating two different test objects of varying usability. We decided to use the two websites http://www.conrad.de and http://www.pearl.de, which are two comparable German e-shop websites of two major German electronics retailers. We conducted a heuristic expert evaluation via Nielsen's 10 Usability Heuristics [29] and were able to identify multiple usability issues of higher severity for pearl.de and very few for conrad.de. While conrad.de has an overall clear structure and modern design, pearl.de lacks modern functions like filter possibilities and offers a more cluttered and complex design. Furthermore, we were able to design tasks that are overall similar in complexity and content and the same concerning the formulation.

4.2 Tasks

We have designed three tasks and an introductory exploration that can be solved on both websites. However, based on the results of the heuristic expert evaluation, we purposely tried to integrate usability issues for pearl.de. While this step certainly skews the usability metrics, please note that our main goal is not to identify and compare the usability between the two websites, but to induce variance in participant behavior.

First, participants were advised to shortly explore the website and report upon their first impressions. We will also refer to this exploration part as task 0. As first real task, participants had to search for an induction charger for smartphones with specific technical attributes. Participants solved this task by adding three possible induction chargers into the shopping basket (task 1). In the second task participants had to interact with the shopping task and reduce the selection just to one device according to a technical attribute (task 2). The third task was an information seeking task: Participants had to find out how to return received devices. They were instructed to find the return form of the e-shop (task 3). All three tasks were solvable with similar steps for both shops.

4.3 Procedure

The entire usability test was performed in a room on a laptop. Participants were introduced into the test procedure by a coordinator and signed a consent form that we are allowed to record the test and participants' face via the webcam. The tasks were introduced by the coordinator, and participants were also offered a sheet of paper with the tasks to read them. The entire test was directed by the test coordinator who was available for questions and observed task completion. Participants were introduced into the concept of thinking aloud and advised to perform the test according to this method. We recorded audio, screen and the face of every participant with the free recording software *Open Broadcaster Software* (OBS).[3] The video files were exported in mp4-format. After all three tasks participants filled out a demographic questionnaire and the *System Usability Scale*.

4.4 Usability Metrics

As major performance metric we analyze the *task completion rate*. This is an established metric in usability testing [2]. The task completion was defined as binary metric per task, meaning a task could either be fully completed or not. The test coordinator observed task outcome and identified the task completion. Tasks could be failed if partici-

3 https://obsproject.com/de

pants informed the coordinator orally that they quit and are not able to complete the task. The overall task completion rate is the percentage of all successfully completed task among the number of all tasks. We did not include the first exploration task into this calculation because this task could not be failed.

Another established usability metric we used is the System Usability Scale (SUS). The questionnaire consists of 10 statements about usability. Participants rated those statements on a five point Likert-scale. The result of these ratings is a metric for usability on a scale from 0 (very low usability) to 100 (very high usability). We used a German version of the SUS [33].

We also recorded the task completion time. However, since we applied thinking aloud, this measure is skewed by the individual communicativeness of each participant. Therefore, we will not include this metric in our analysis but will report descriptive statistics and important findings to support the interpretation and discussion.

4.5 Sentiment/Emotion Metrics

As dependent variables, we examine various tools for measuring sentiment and emotion for various modalities. We decided to use general purpose tools and approaches that are not optimized for the specific task of usability testing since the usage of accessible, free and easy-to-apply methods and tools is certainly more likely for usability engineering in research and industry. Furthermore, we focus on the concepts of valence, meaning if the emotions expressed are overall rather positive or negative. This facilitates the interpretation and comparison of the various channels, since the emotion definitions and representations vary vastly by the modality. However, we will investigate the individual emotions of the audio and face analysis separately via regression models.

4.5.1 Text-Based Sentiment Analysis

We perform text-based sentiment analysis on text received via speech-to-text software on the videos of the post; thus we are applying text-based sentiment analysis on what the participants were saying during the tests. The transformation to text was done via the *Google Cloud Speech API*[4] which is considered state-of-the-art for speech-to-text tasks. Please note, however, that we did not precisely evaluate the performance on our videos. Random testing showed that the API performs rather well. Nevertheless, we did identify problems when participants had a strong dialect, accent or talked unclearly. We did however remove such extreme cases. For sentiment analysis we applied a lexicon-based approach by using the German general-purpose lexicon *SentiWS* [34]. This is an established sentiment lexicon and has been proven beneficial compared to other lexicons in some areas [39]. It basically consists of over 30,000 terms (including inflections of words) that are annotated on a scale from −3 to 3 concerning the textual valence (whether the term is rather negatively or positively connoted). By summing up the values of all terms apparent in a text one can acquire an overall valence value. We did so for all the texts received by the Google Cloud Speech API.

4.5.2 Speech/Audio-Based Emotion Analysis

To perform audio-based emotion analysis we first extracted the audio channel from the video file in 16 bit PCM wave-format. We then cut each audio file into files of ten second length since our speech emotion analysis software works better with shorter audio snippets. For the emotion analysis we use the free developer version of the software *Vokaturi*.[5] *Vokaturi* is an emotion recognition tool for spoken language using machine learning techniques. It is recommended as the best free general purpose software for emotion analysis of spoken language (cf. Garcia-Garcia et al. [16]). *Vokaturi* uses machine learning on two larger databases with voice- and audio-based features. One can enter an audio-file to receive numerical values on a range from 0 (none) to 1 (a lot) for the five categories neutrality, fear, sadness, anger and happiness. The values represent probabilities. We entered the 10-second snippets of all tasks to receive an overall value for all categories. To map this output to the binary concept of valence we decided to employ a heuristic approach: We sum up the values for negativity (fear, sadness and anger) and deduct this value from happiness. We will refer to this value as audio valence. In the following analysis we will also investigate the role of neutrality.

4.5.3 Video/Face-Based Emotion Analysis

Finally, to perform the emotion analysis based on the faces of the participants recorded via the webcam, we use the tool *OpenFace 2.2.0* [3]. This is an open source face analysis software and serves as a practical solution in various research areas (e. g. Fydanaki & Geradts [15]; Santoso &

4 https://cloud.google.com/speech-to-text/docs/apis

5 https://vokaturi.com

Kusuma [36]). Similar to other face recognition software, *OpenFace* deconstructs facial expression according to the *Facial Action Coding System* (FACS). This is an established concept to measure facial expressions derived from psychology (cf. Sayette et al. [37]). Facial action units are small face movements that are performed unconsciously and hard to control. Examples for an action unit are the cheek raiser or the lip corner puller, which are both signs for happiness. *OpenFace* offers the action unit results for a subset of the FACS and marks for every frame if the action unit is apparent and to which intensity (on 5-point scale). We sum up the values for all relevant action units for every task and overall. Ekman and Friesen [14] define calculations based on this action units to acquire values for various emotional expressions. We performed these calculations to acquire values for the emotion categories happiness, sadness, surprise, anger, fear and disgust. Similar to the audio approach we applied a heuristic to gather an overall valence value by summing up the negative emotion values (anger, fear, disgust, sadness) and deducting this value from happiness. We neglect *surprise* for these calculations since it as an ambivalent emotion but will analyze it as single value later on.

4.6 Sample

We performed the usability test with 125 participants, 59 female and 66 male. The average age is 28.48, the age ranges between 15 and 59 years. The original number of participants was higher but we eliminated participants that were familiar with the websites and test recordings with obvious technical problems that disallowed the usage of our software. To acquire such a large number of participants we compromised the control over the experiment to a minor extent: The tests were performed in different locations and by different test coordinators. The introduction and overall setting remained the same.

5 Results

We will first report upon descriptive statistics on the independent and dependent variables and then continue with the statistical analysis of the research question.

5.1 Descriptive Statistics

5.1.1 Usability Metrics

Table 1 summarizes the results for the SUS scores.

Table 1: SUS results.

Shop	Min	Avg	Med	Max	Std
conrad.de	20	72.56	75	100	19.32
pearl.de	0	26.32	25	87.5	17.57

As expected and intended, the shop pearl.de scores much lower on average (M = 26.32) than conrad.de (M = 72.56). The lower level of usability is also shown by regarding the task completion rate for each task. Table 2 illustrates the percentage of participants that successfully completed the task. While 80 % of all tasks were successfully completed for conrad.de, this number is much lower for pearl.de (44.7 %).

Table 2: Task completion rate per shop and task.

Shop	Task	Completion Rate
conrad.de	1	78.4 %
	2	80 %
	3	80.8 %
	total	79.7 %
pearl.de	1	40.8 %
	2	53.6 %
	3	36.8 %
	total	44.7 %

The analysis of the task completion time also shows that participants took more time on average and for most tasks on pearl.de (Table 3).

Table 3: Task completion time per shop and task (in seconds).

Shop	Task	Min	Avg	Med	Max	Std
conrad.de	0	6	82.71	45	2160	202.58
	1	92	444.73	396	1460	242.28
	2	23	134.07	109	752	100.79
	3	33	145.66	114	499	99.40
	total	194	809.22	759	2438	368.65
pearl.de	0	2	67.86	49	579	67.514
	1	130	575.59	534	1617	267.22
	2	22	79.75	73.50	508	55.66
	3	40	249.49	225	831	151.77
	total	279	964.23	919	2366	360.34

Please note that test persons did perform thinking aloud during the tests, therefore the time is only of limited use as a performance metric. Nevertheless, the descriptive statistics showed that the participants took around 13.5 minutes (809.22 seconds) to complete the usability test with conrad.de and around 16 minutes with pearl.de. On

average, participants completed the usability test for both shops in around 30 minutes (1773.45 seconds).

5.1.2 Sentiment/Emotion Metrics

The audio-based statistics are illustrated in Table 4.

Table 4: Descriptive statistics for audio-based emotion analysis.

Shop	Category	Min	Avg	Max	Std
conrad.de	Neutrality	0.00	1.22	3.82	1.11
	Happiness	0.00	.51	2.72	.663
	Sadness	.11	.905	2.96	.565
	Anger	0.00	.168	1.33	.242
	Fear	0.00	.375	2.29	.467
pearl.de	Neutrality	0.00	1.16	3.38	1.02
	Happiness	0.00	.495	2.40	.646
	Sadness	.079	.89	3.01	.58
	Anger	0.00	.191	1.71	.303
	Fear	0.00	.367	2.31	.482

Note that the 0.00-values represent very low values that were rounded to 0.00. We have no participant where emotions were not detected at all, but we do have participants with very low levels of emotions (The same holds true for Table 5). Overall, the descriptive results are ambivalent concerning expectations. Neutrality is higher on conrad.de as well as happiness. However, some results are counterintuitive as fear and sadness are higher for *Conrad* although this being the better rated shop. Nevertheless, all results are very close to each other, so overall this approach does not find strong differences between the shops.

Table 5 describes the descriptive statistics for the facial emotion recognition.

Table 5: Descriptive statistics for video-based emotion analysis.

Shop	Category	Min	Avg	Max	Std
conrad.de	Happiness	0.00	.128	.635	.146
	Sadness	.024	.287	1.59	.294
	Surprise	.048	.246	.541	.103
	Fear	.061	.575	1.82	.374
	Anger	.030	.338	1.63	.356
	Disgust	.004	.078	.257	.050
pearl.de	Happiness	0.00	.132	.727	.149
	Sadness	.033	.302	1.73	.291
	Surprise	.043	.237	.750	.111
	Fear	.074	.578	1.99	.378
	Anger	.032	.359	1.95	.364
	Disgust	0.00	0.00	.229	0.00

Similar to the audio-based subcategories, the overall averages are rather similar and ambivalent. Happiness is higher for pearl.de while disgust is higher for conrad.de, which appears to be counterintuitive.

Table 6 summarizes the data for the textual sentiment analysis and the valence heuristics we derived from the subcategories of the other modalities.

Table 6: Descriptive statistics for textual sentiment analysis and valence values.

Shop	Category	Min	Avg	Max	Std
conrad.de	Text	−4.06	.109	4.19	1.51
	Audio Valence	−3.13	−.938	1.82	.889
	Video Valence	−4.95	−1.15	−.116	.980
pearl.de	Text	−5.64	−.207	4.12	1.73
	Audio Valence	−3.09	−.950	1.63	.912
	Video Valence	−5.38	−1.18	−.119	.985

The only clear and consistent difference can be found for the textual sentiment analysis for which conrad.de shows a positive average while pearl.de is negative. Please note that the tendency towards negativity for the audio- and video-based approaches might very well be due to the way we calculate this heuristic. Thus, an interpretation of the absolute values is difficult, however comparisons are still valid. The results show marginal differences but in contrast to the subcategories are as expected in the sense that pearl.de has lower averages than conrad.de.

5.2 Inference Statistics

In the following we examine the results of our research questions.

5.2.1 RQ1: Correlations (Usability Metrics – Sentiment/Emotion Metrics)

We test this relationship on the set of all single usability tests we performed resulting in N = 250. We will solely examine the accumulated values over all tasks. Our variables are not equally distributed. Therefore, we use the *Kendall Tau rank correlation coefficient* to analyze the relationship. This metric is recommended for data that is not normally distributed and the metric is rather robust and conservative [9]. All preconditions are satisfied by the data. We will solely report upon results that show a statistically significant correlation. As significance level we choose p = .05.

Text-Based Sentiment Analysis
We identified a significant positive correlation between the SUS-values and the result of the textual sentiment analysis summing up the results for all tasks; the higher the valence value the higher the score. However the correlation is very weak (τ = .086; p = .046). There is no significant correlation with the task completion rate.

Audio/Speech Based Emotion Analysis
For the SUS score two of the emotion categories of the speech based emotion analysis showed a significant correlation: (1) the category neutrality correlates positively with the SUS score. The correlation is weak (τ = .089; p = .041). The more neutral the speech expression, the higher the score. (2) Happiness correlates negatively and weak with the SUS score (τ = −.089; p = .040). This correlation is counterintuitive since the SUS score is higher the lower the happiness values. Note that the valence heuristic we calculated for this modality did not show any significant correlations.

Video/Face Based Emotion Analysis
None of the emotion categories showed a significant correlation with the SUS score or the task completion rate. Similar to the audio approach, the heuristic for valence does not correlate significantly with the SUS score or the task completion rate.

5.2.2 RQ2: Group Based Comparisons – Differences Between the Websites

We investigated if the various sentiment and emotion metrics of the applied tools can be used to differentiate the tested shops similar to the usability metrics. Since every participant did test each web shop in counterbalanced manner, we regard our data as dependent samples. We applied the *Wilcoxon signed rank test* for dependent samples. This test is not dependent on normal distribution and all other prerequisites for this test are met by our data.

As the descriptive data already indicate, all usability metrics show a significant difference between both web shops with conrad.de being higher rated concerning the SUS score (Z = −9.435; p < .000; M(conrad.de) = 72.56; M(pearl.de) = 26.32) and showing higher levels of task completion rate (Z = −8.080; p < .000; M(conrad.de) = .933; M(pearl.de) = .530). On a minor note: this difference holds also true for the task completion time (Z = −4.558; p < .000; M(conrad.de) = 809.22; M(pearl.de) = 964.23).

However, the Wilcoxon signed rank test did not show any significant group differences of the web shops concerning the sentiment and emotion metrics for any of the modalities. Thus, these metrics cannot be used to differentiate web shops of varying usability in our study.

5.2.3 RQ3: Regression Analysis

We performed several linear regression analysis. For one, we wanted to investigate how and to what extent the subcategories of the modalities contribute to the prediction of usability metrics and if combinations of subcategories can explain a significant amount of variance as a combined model instead of a single factor. Secondly, we examine if the combination of the modalities can successfully predict usability metrics as regression models; thus we test if a multimodal approach based on our heuristics for valence performs better than the prediction via individual factors. In the following, we will first report if the linear regression model shows a significant effect. We will also report the standardized regression coefficients of all variables of a model to analyze if and how the individual factors influence the model.

Audio/Speech Based Emotion Analysis
The model consisting of all five speech based emotion categories shows no significant linear regression with the SUS score as independent variable (F(5, 245) = 1.675, p = .141). Table 7 illustrates the effect of the individual regression coefficients.

Table 7: Regression model results for speech based emotion analysis subcategories and SUS.

Model	Standardized Regression Coefficient	Sig.
Audio_Neutrality	.153	.112
Audio_Happiness	.059	.526
Audio_Sadness	.139	.060
Audio_Anger	−.002	.976
Audio_Fear	−.045	.569

Similar to the correlation analysis, neutrality is the strongest predictor although not significant. It is notable that all coefficients show a direction as expected except for sadness which actually shows a positive relationship, meaning the higher this value the higher the SUS score. However this relationship is not significant.

The audio-based-model is also not significant for the task completion rate (F(5, 245), p = .449). Neutrality is the strongest regression coefficient showing a significant positive relationship. Happiness is the next strongest co-

efficient, though not significant. Fear and sadness show a weak but counterintuitive negative correlation (see Table 8).

Table 8: Regression model results for speech based emotion analysis subcategories and task completion.

Model	Standardized Regression Coefficient	Sig.
Audio_Neutrality	.197	.042
Audio_Happiness	.132	.161
Audio_Sadness	.038	.611
Audio_Anger	−.019	.769
Audio_Fear	.048	.549

Video/Face Based Emotion Analysis

The model consisting of all six factors of the facial emotion analysis cannot be proven as a significant linear regression with the SUS score ($F(6, 246) = .318$; $p = .927$).

Table 9: Regression model results for facial emotion analysis subcategories and SUS.

Model	Standardized Regression Coefficient	Sig.
Video_Happiness	−.059	.400
Video_Sadness	−.007	.953
Video_Surprise	−.135	.505
Video_Fear	.386	.572
Video_Anger	−.392	.543
Video_Disgust	. 049	.500

The individual analysis of the emotion categories shows no significant relationship with the SUS score. The strongest individual factors are fear and anger with fear having a positive relationship though being a negative emotion type (see Table 9). The video-based-model shows also no significant linear regression towards the task completion rate ($F(6, 246) = 1.506$; $p = .177$).

Similar to the regression model for SUS scores, fear and anger show the strongest coefficients with fear being positive and anger negative. However, none of the factors shows a significant relationship (see Table 10).

Multimodal Regression Model

We examined if a regression model consisting of variables of all three modalities can successfully predict SUS scores or the task completion rate. The model we create consists of the three valence scores of the three modalities.

Table 10: Regression model results for facial emotion analysis subcategories and task completion rate.

Model	Standardized Regression Coefficient	Sig.
Video_Happiness	.025	.712
Video_Sadness	−.160	.187
Video_Surprise	.050	.801
Video_Fear	.252	.707
Video_Anger	−.229	.718
Video_Disgust	.034	.634

Considering the SUS value this model shows no significant linear regression model ($F(3, 244) = 2.424$; $p = .066$). Regarding the individual factors however, the textual sentiment analysis shows to be the strongest predictor and is significant in itself. The audio valence shows a negative relationship which is counterintuitive since this means the higher the SUS scores the lower the valence measured by the audio emotion analysis (Table 11).

Table 11: Regression model results for valence metrics and SUS.

Model	Standardized Regression Coefficient	Sig.
Text_Valence	.155	.017
Audio_Valence	−.100	.121
Video_Valence	.010	.873

The multimodal regression model shows no significant relationship for the task completion rate ($F(3, 244) = 1.765$; $p = .154$). The strongest regression coefficient is the video valence, though again not significant as individual factor (see Table 12).

Table 12: Regression model results for valence metrics and SUS.

Model	Standardized Regression Coefficient	Sig.
Text_Valence	.071	.273
Audio_Valence	.028	.663
Video_Valence	.116	.070

6 Discussion

In the following section, we want to interpret the performance of our tools for our specific use case and discuss the role and limitations of emotion recognition software in us-

ability engineering. Finally, we make some suggestions for how to continue research in this field.

6.1 Tool Performance

Summing up, our study must be regarded as inconclusive. The majority of metrics that our tools produce show no significant relationships or are even counterintuitive with weak correlation values. The textual sentiment analysis shows a weak positive correlation with the SUS ratings. The more positive words are used the better the SUS score. We argue that the reason might be that the conceptual idea behind textual sentiment analysis is closer to what SUS scores measure. The concept of the textual sentiment analysis is to predict the opinion or sentiment towards something [23, p. 1], similar like the SUS score measures opinions of users towards a system. Therefore, sentiment lexicons contain a lot of words fitting for describing products as positive or negative. On the other hand, the speech and facial emotion recognition systems try to predict how humans feel themselves which might be a concept that is less in line with regular usability metrics. Another significant metric was the proportion of neutrality in speech-based emotion analysis. The more neutral the tone of the voice of users during the test, the better the usability of the website in our study, which is not the case for happiness. This might be an indicator that participants do not tend to be overly joyous during standardized usability tests but rather that calmness and missing emotionality point to a satisfying experience. Other than that, the audio-based approach also showed counterintuitive results. A reason for that might be that the setting of usability tests is quite different from the usual application area of these tools: telephone counselling.

We got the impression that the facial emotion recognition did not work as we expected it at all, which is surprising since, in general, facial emotion recognition is supposed to perform better than speech emotion recognition [19] and achieves accuracies up to 98 % percent in other studies (cf. Saxena et al. [38]). Some counterintuitive results (e. g. the correlation direction of happiness) lead us to the assumption that certain facial expressions for certain situations in usability tests might not resemble what they usually mean. These inconclusive findings are in line with previous research: Thüring and Mahlke [46] did not find any significant correlations with facial emotion recognition and SUS scores. Xu et al. [48] did find significant differences for the arousal value of facial emotion recognition when comparing two user interfaces, however, they did not find significant differences for the valence value.

We investigated our data more in-depth with various approaches. Instead of analyzing the entire test, we split the analysis into the subtasks but the results remained the same. We also analyzed the scatter plots of our correlations to examine if the correlation might not be linear but follows some other shape that our statistical tests do not verify. However, we could not find any specific shape being it inverted-U or any other form. Therefore, we come to the following conclusion: Although we found some punctual significant correlations, they are rather weak which leads us to state that our study bears inconclusive or even negative results. Based on our results and the majority of previous research, emotion recognition software is not a stable and consistent tool to predict usability.

However, these tools might still be beneficial for other tasks and goals than just the one we investigated. Johannsson et al. [20] use facial emotion recognition software to investigate important sequences during a test. For example: they look at strong manifestations of negative emotions to detect usability issues. We recommend researchers and usability practitioners to also further work towards this direction.

6.2 Emotion Recognition and Usability Engineering

There are two explanations for our results: (1) emotion recognition tools have inherent limitations connected to the specific task of usability tests and (2) the used tools and methods are not sophisticated enough or do not fit this specific task. Considering (1): some problems with text might be that the amount of text to be included in the analysis is rather limited. On average the thinking aloud scripts consist of 100 words (around 10 sentences), but sentiment analysis has been shown to perform better on larger text documents [23, p. 47]. We recommend to include test coordinator protocols, open-ended questions at the end of the test and other material to increase the amount of text. For audio, the setting of a usability test might just include too many disruptive factors: The fact that sometimes the test coordinator did also speak during the tests and that participants don't focus on speaking clearly and loudly might have skewed the audio emotion analysis. Furthermore, the emotional expressions during controlled usability tests are in general less expressive than in real life. Participants usually try to suppress strong emotional outbursts and tend to behave calm even when confronted with problems

or joy. This might also influence the facial emotion recognition. In addition, especially in our setting, the length of a usability test as well as the factor that users had to talk constantly due to the thinking aloud method leads to problems in correct facial emotion recognition. The large variety of facial expressions that can be found in such a lengthy task-based setting while speaking is quite contrary to the static frames most of this software is trained for. Talking might lead to facial expressions that are falsely classified. This can of course be easily verified by performing tests without thinking aloud.

Nevertheless, we see the most value in the exploration of more advanced emotion recognition approaches adjusted to our task. For this study, we used general-purpose tools; the trend in the research area is however to adjust and fine-tune methods to a specific use case if the use case is very unusual [38]. This is obviously the case for usability tests and the results show that the tools are trained and designed for other use cases. The current state-of-the-art in text based sentiment analysis is the usage of pretrained word embeddings [49] like BERT [12] and fine-tuning them to a specific domain via deep neural networks [42]. This process is dependent on large amounts of training material, meaning texts of usability tests that are annotated concerning their valence. The same holds true for audio emotion analysis which is dependent of training material for deep learning approaches, as well. Facial emotion recognition offers a plethora of machine learning techniques and standardized training corpora (cf. Saxena et al. [38]). However, when analyzing the training images it becomes clear that they are far apart from the setting and the facial expressions of a usability test. The combination of all channels via multimodal emotion recognition [32] might be a beneficial way to avoid the singular limitations.

Independent of the channel, to optimize emotion analysis for usability engineering by gathering a lot of annotated data, a joint effort of the HCI-community is necessary. It is important for researchers to start reflecting about possibilities to publish recorded tests and supplementary material with usability results as corpora and annotate this data, a process well-established in the AI-community. While there is certainly a problem concerning information privacy, even a small fraction of annotated test results acquired in the community during a year would enable the stable application of machine learning techniques and emotion recognition adjusted to this use case. We plan to perform such annotation studies on fractions of our acquired data to advance this research idea.

References

[1] Agarwal, A., & Meyer, A. (2009). Beyond usability: evaluating emotional response as an integral part of the user experience. In CHI'09 Extended Abstracts on Human Factors in Computing Systems (pp. 2919–2930).

[2] Albert, W., & Tullis, T. (2013). Measuring the user experience: collecting, analyzing, and presenting usability metrics. Newnes.

[3] Baltrušaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018, May). Openface 2.0: Facial behavior analysis toolkit. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 59–66). IEEE.

[4] Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry, 25(1), 49–59.

[5] Brave, S., & Nass, C. (2002). Emotion in human-computer interaction. In The human-computer interaction handbook (pp. 103–118). CRC Press.

[6] Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4–7.

[7] Bruun, A., Law, E. L. C., Heintz, M., & Eriksen, P. S. (2016, October). Asserting real-time emotions through cued-recall: Is it valid? In Proceedings of the 9th Nordic Conference on Human-Computer Interaction (pp. 1–10).

[8] Champney, R. K., & Stanney, K. M. (2007, October). Using emotions in usability. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 51(17), 1044–1049. Sage CA: Los Angeles, CA: SAGE Publications.

[9] Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. Statistical methods & applications, 19(4), 497–515.

[10] Daily, S. B., James, M. T., Cherry, D., Porter III, J. J., Darnell, S. S., Isaac, J., & Roy, T. (2017). Affective computing: historical foundations, current applications, and future trends. In Emotions and Affect in Human Factors and Human-Computer Interaction (pp. 213–231). Academic Press.

[11] Desmet, P. M. A. (2000). Emotion through expression: Designing mobile telephones with An emotional fit. In Report of modeling the evaluation structure of Kansei.

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technology Conference (NAACL-HLT 2019) (pp. 4171–4186). ACL.

[13] Dubey, S. K., & Rana, A. (2010). Analytical roadmap to usability definitions and decompositions. International Journal of Engineering Science and Technology, 2(9), 4723–4729.

[14] Ekman, P., & Friesen, W. V. (1978). Facial action coding system: Investigator's guide. Consulting Psychologists Press.

[15] Fydanaki, A., & Geradts, Z. (2018). Evaluating OpenFace: an open-source automatic facial comparison algorithm for forensics. Forensic sciences research, 3(3), 202–209.

[16] Garcia-Garcia, J. M., Penichet, V. M., & Lozano, M. D. (2017, September). Emotion detection: a technology review. In Proceedings of the XVIII International Conference on Human Computer Interaction (pp. 1–8).

[17] Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In Mensch & computer 2003 (pp. 187–196). Vieweg+Teubner Verlag.
[18] Hassenzahl, M., & Ullrich, D. (2007). To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. Interacting with computers, 19(4), 429–437.
[19] Hudlicka, E. (2003). To feel or not to feel: The role of affect in human-computer interaction. International journal of human-computer studies, 59(1-2), 1–32.
[20] Johanssen, J. O., Bernius, J. P., & Bruegge, B. (2019, May). Toward usability problem identification based on user emotions derived from facial expressions. In 2019 IEEE/ACM 4th International Workshop on Emotion Awareness in Software Engineering (SEmotion) (pp. 1–7). IEEE.
[21] Kim, E., & Klinger, R. (2018). A survey on sentiment and emotion analysis for computational literary studies. arXiv preprint arXiv:1808.03137.
[22] Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. Motivation and emotion, 5(4), 345–379.
[23] Liu, B. (2016). Sentiment Analysis. Mining Opinions, Sentiments and Emotions. New York: Cambridge University Press.
[24] Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis – A review of research topics, venues, and top cited papers. Computer Science Review, 27, 16–32.
[25] McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R. E. (2016, May). AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. In Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems (pp. 3723–3726).
[26] Mehrabian, A., & Russell, J. A. (1974). An approach to environmental psychology. the MIT Press.
[27] Mulligan, K., & Scherer, K. R. (2012). Toward a working definition of emotion. Emotion Review, 4(4), 345–357.
[28] Nielsen, J. (1994). Guerrilla HCI: Using discount usability engineering to penetrate the intimidation barrier. In Cost-justifying usability (pp. 245–272).
[29] Nielsen, J. (1995). 10 usability heuristics for user interface design. Nielsen Norman Group, 1(1).
[30] Picard, R. W. (2000). Affective computing. MIT press.
[31] Plutchik, R. (1980). Emotion: A psychoevolutionary synthesis. New York: Harper & Row.
[32] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98–125.
[33] Rauer, M. (2011). Quantitative usability-analysen mit der system usability scale (SUS). Seibert Media. Retrieved from https://blog.seibert-media.net/blog/2011/04/11/usablility-analysen-system-usability-scale-sus/, zuletzt geprüft am, 19, 2016.
[34] Remus, R., Quasthoff, U., & Heyer, G. (2010, May). SentiWS – A Publicly Available German-language Resource for Sentiment Analysis. In LREC.
[35] Russell, J. A. (1980). A circumplex model of affect. Journal of personality and social psychology, 39(6), 1161.
[36] Santoso, K., & Kusuma, G. P. (2018). Face recognition using modified OpenFace. Procedia Computer Science, 135, 510–517.
[37] Sayette, M. A., Cohn, J. F., Wertz, J. M., Perrott, M. A., & Parrott, D. J. (2001). A psychometric evaluation of the facial action coding system for assessing spontaneous expression. Journal of Nonverbal Behavior, 25(3), 167–185.
[38] Saxena, A., Khanna, A., & Gupta, D. (2020). Emotion recognition and detection methods: A comprehensive survey. Journal of Artificial Intelligence and Systems, 2(1), 53–79.
[39] Schmidt, T., & Burghardt, M. (2018). An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In: Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (pp. 139–149). Association for Computational Linguistics, Santa Fe, New Mexico.
[40] Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Construction of a Benchmark for the User Experience Questionnaire (UEQ). IJIMAI, 4(4), 40–44.
[41] Schröder, M., Pirker, H., & Lamolle, M. (2006, May). First suggestions for an emotion annotation and representation language. In Proceedings of LREC (Vol. 6, pp. 88–92).
[42] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, October). How to fine-tune bert for text classification? In China National Conference on Chinese Computational Linguistics (pp. 194–206). Springer, Cham.
[43] Suzuki, K., Camurri, A., Ferrentino, P., & Hashimoto, S. (1998, October). Intelligent agent system for human-robot interaction through artificial emotion. In SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 98CH36218) (Vol. 2, pp. 1055–1060). IEEE.
[44] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267–307.
[45] Tao, J., & Tan, T. (2005, October). Affective computing: A review. In International Conference on Affective computing and intelligent interaction (pp. 981–995). Springer, Berlin, Heidelberg.
[46] Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human-technology interaction. International journal of psychology, 42(4), 253–264.
[47] Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. International Journal of Advanced Research in Computer Science and Soft-ware Engineering, 2(6), 282–292.
[48] Xu, N., Guo, G., Lai, H., & Chen, H. (2018). Usability Study of Two In-Vehicle Information Systems Using Finger Tracking and Facial Expression Recognition Technology. International Journal of Human-Computer Interaction, 34(11), 1032–1044.
[49] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), https://doi.org/10.1002/widm.1253.

Bionotes

Thomas Schmidt
Media Informatics Group, University of Regensburg, Regensburg, Germany
thomas.schmidt@ur.de

Thomas Schmidt is a PhD student, research assistant and lecturer at the chair for media informatics, University of Regensburg. Among his research interests are Digital Humanities, text mining, human-computer interaction (HCI) and information behavior. In his previous HCI endeavors, he investigated the influence of gender and personality on user behavior, the application of affective computing in various interaction settings and the effects of UI aesthetics. He currently works in the computational literary studies project "Emotions in Drama" investigating sentiment and emotions in historic German plays via computational methods. He is also responsible for the coordination of the master degree Digital Humanities at the University of Regensburg. In his PhD, he explores the application of multimodal emotion analysis in movies and theater recordings.

Miriam Schlindwein
Media Informatics Group, University of Regensburg, Regensburg, Germany
miriam.schlindwein@stud.uni-regensburg.de

Miriam Schlindwein is a master student of media informatics at the University of Regensburg and holds a bachelor of arts in media informatics and German language and literature. Among her research interests are usability engineering, human-computer interaction and Digital Humanities. In her master thesis, she investigates user-centered process support for counting of local government elections in Bavaria.

Katharina Lichtner
Media Informatics Group, University of Regensburg, Regensburg, Germany
katharina.lichtner@stud.uni-regensburg.de

Katharina Lichtner is a master student of media informatics at the University of Regensburg and holds a bachelor of arts in media informatics and information science. Among her research interests are usability engineering, human-computer-interaction and Digital Humanities. In her master thesis, she creates and evaluates a concept for interaction with information formats on large-area displays.

Christian Wolff
Media Informatics Group, University of Regensburg, Regensburg, Germany
christian.wolff@ur.de

Prof. Dr. Christian Wolff has been professor for media informatics at the University of Regensburg since 2003. After his Ph. D. thesis, in which he designed an interactive retrieval frontend for factual data (1994, University of Regensburg), he worked as an assistant professor at the Computer Science Department of the University of Leipzig from 1994 to 2001 and became professor for media informatics at the Chemnitz University of Technology in 2002. He has a long record of research in electronic publishing, information retrieval and HCI. Currently, there is a strong focus on new interaction technologies like eye tracking or virtual reality in his group. In 2019, he has been appointed to the founding commission for the new faculty for informatics and data science in Regensburg where is acting dean of studies.