User Modeling and User-Adapted Interaction (UMUAI) manuscript No.
(will be inserted by the editor)

Extracting Relevance and Affect Information from Physiological Text Annotation

Oswald Barral · Ilkka Kosunen · Tuukka Ruotsalo · Michiel M. Spapé · Manuel J. A. Eugster · Niklas Ravaja · Samuel Kaski · Giulio Jacucci

Received: date / Accepted: date

Abstract We present physiological text annotation, which refers to the practice of associating physiological responses to text content in order to infer characteristics of the user information needs and affective responses. Text annotation is a laborious task, and implicit feedback has been studied as a way to collect annotations without requiring any explicit action from the user. Previous work has explored behavioral signals, such as clicks or dwell time to automatically infer annotations, and physiological signals have mostly been explored for image or video content. We report on two experiments in which physiological text annotation is studied first to 1) indicate perceived relevance and then to 2) indicate affective responses of the users. The first experiment tackles the user's perception of relevance of an information item, which is fundamental towards revealing the user's information needs. The second experiment is then aimed at revealing the user's affective responses towards a -relevant- text document. Results show that physiological user signals are associated with relevance and affect. In particular, electrodermal activity (EDA) was found to be different when users read relevant content than when they read irrelevant content and was found to be lower when reading texts with negative emotional content than when reading texts

This paper or a similar version is not currently under review by a journal or conference, nor will it be submitted to such within the next three months. This paper is void of plagiarism or self-plagiarism as defined in Section 1 of ACM's Policy and Procedures on Plagiarism.

Oswald Barral, Ilkka Kosunen, Giulio Jacucci
Helsinki Institute for Information Technology HIIT
University of Helsinki, Department of Computer Science, PO Box 68, 00014, Finland
E-mail: first.last@hiit.fi

Tuukka Ruotsalo, Michiel M. Spapé, Niklas Ravaja, Giulio Jacucci, Manuel J.A. Eugster, Samuel Kaski
Helsinki Institute for Information Technology HIIT
Aalto University, Department of Computer Science, PO Box 15400, 00076, Finland
E-mail: first.last@hiit.fi

Niklas Ravaja
Helsinki Institute for Information Technology HIIT
University of Helsinki, Department of Social Research, PO Box 54, 00014, Finland
E-mail: first.last@hiit.fi

with neutral content. Together, the experiments show that physiological text annotation can provide valuable implicit inputs for personalized systems. We discuss how our findings help design personalized systems that can annotate digital content using human physiology without the need for any explicit user interaction.

Keywords Physiological annotation · Electrodermal activity · Peripheral physiology · Perceived relevance · Affective feedback · Text annotation

Acknowledgements This work has been partly supported by the Academy of Finland (278090, Multivire, 255725; and the Finnish Centre of Excellence in Computational Inference Research COIN) and MindSee (FP7 - ICT; Grant Agreement # 611570). Certain data included herein are derived from the Web of Science prepared by THOMSON REUTERS, Inc., Philadelphia, Pennsylvania, USA: Copyright THOMSON REUTERS, 2011. All rights reserved. Data are also included from the Digital Libraries of the ACM, IEEE, and Springer.

1 Introduction

The proliferation of Web content has allowed users to access, consume and annotate a large variety of media content online. Annotation refers to adding descriptive metadata to digital content and has traditionally been a critical backbone of many digital media services. It allows content management and analysis by enabling additional detailed information about the content. On an individual level, users can benefit from annotation since it can enhance content delivery, personalization, and recommendation through analysis of additional metadata that capture user preferences (Adomavicius and Tuzhilin, 2005; Herlocker et al., 2004).

The most common practice to annotate content is by explicit user-system interaction, i.e., users are asked to manually provide feedback on the content. Current feedback mechanisms include bookmarking, rating, and tagging. To a certain level, these are digital adaptations of traditional pen-based annotations. However, explicit feedback can be tedious and users usually annotate content only if they have a clear incentive to do so. It has been shown that users are not always willing to interrupt their task to provide feedback even when they are aware that this could lead to benefits in the subsequent interactions with the system (Kelly and Fu, 2006).

Another way to annotate content is by implicitly monitoring the users. This type of annotation, which does not rely on direct input from the user, has the benefit that it requires no effort from the user and does not require interrupting the workflow. Behavioral signals that have been used for this purpose range from measuring dwell time to monitoring clicks or scrolling behavior (Kelly and Teevan, 2003; Soleymani and Pantic, 2012). For instance, YouTube recommendations are not only based on the users' likes or subscriptions to channels, but also take into account the time users spend watching videos (Davidson et al., 2010).

The popularity of using implicitly generated metadata is increasing. Studies have shown that although these measures can have high throughput, they can fail to capture the actual level of user-engagement with the content. For example, a user could be playing a YouTube video but not paying attention to it (Davidson et al., 2010). Furthermore, these measures fail to capture the emotional state of the user, which gives important clues to the users' preferences (Tkalcic et al., 2011).

A novel technique for implicit monitoring is annotating content using physiological recordings (Moshfeghi and Jose, 2013; Eugster et al., 2014; Barral et al., 2015; Treacy Solovey et al., 2015). While most information consumption is in textual format, surprisingly little research has been carried out in this area, as most of the work carried out in this direction considers images and videos as the target (Moshfeghi and Jose, 2013; Arapakis et al., 2009b). Text presents additional challenges, as it is more symbolic and contextualized in nature.

This paper presents physiological text annotation, which refers to the practice of associating physiological measures to text content in order to infer characteristics of the user information needs and affective responses. We study whether physiological annotations change reliably due to relevance judgment or affective responses. Conventional corpus annotation relies on the users' explicit input in order to infer general properties of the dataset, which are often used for indexing or explaining better the data (Leech, 1997). Implicit responses of viewers to the corpus have not been considered, and attaching physiological signals as annotations to content could be used to differentiate the affect related to the content at a population level. For example, observing physiological signals from a group of users looking at a news article and annotating this article with these signals, could provide information whether the users -all of them on average- perceive the article with certain affect (e.g., observing whether there is overall increase in skin conductance level when an article is perceived as relevant).

We study physiological text annotation through two experiments, focusing on: 1) the user's perceived relevance of text items, and 2) the user's affective responses to text items. When engaged in an information seeking task, users rely on the perceived relevance of information items to select the items to consume. Once an item is deemed relevant, it is consumed, possibly eliciting affective responses. In the first experiment we explore physiological annotations and their association with the users' perceived relevance of text items. Then, in the second experiment, we explore physiological annotations and their association with the users' affective responses to -relevant- text items. This paper substantially extends the work of Barral et al. (2015) by studying physiological annotation for affective responses in a realistic news reading task, allowing to present a comprehensive description of physiological text annotation. Figure 1 illustrates the conceptual framework for physiological text annotation addressed in this paper.

Physiological annotation for relevance. The first experiment was designed to study the physiological correlates of relevance during reading of text. The process of judging the relevance of an information item is intrinsically related to the affective state of the user and affective responses to the information items being consumed (Cosijn and Ingwersen, 2000). For instance, when searching for information on the internet, it could happen that a page that has relevant content for the purpose of the search is deemed less relevant due to the frustration evoked by a poor page layout. In order to better study physiological responses to the relevance judgment, we wanted to minimize the role of affect in the relevance judgment process. We consequently designed a relevance assessment task in the context of reading scientific articles. This task and domain was considered fairly emotionally neutral, minimizing possible con-

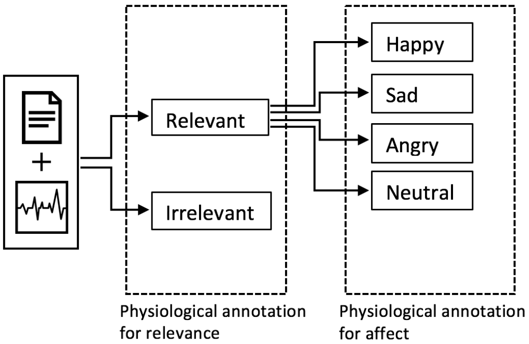

Relevant
Irrelevant
Happy
Sad
Angry
Neutral
Physiological annotation for relevance
Physiological annotation for affect

Fig. 1: Physiological text annotation conceptual framework. First, the user makes a decision on the relevance of a text document. Then, once a document is deemed relevant, it is read, eliciting several affective responses. In a first experiment we investigate reliable changes in the physiology due to the judging the relevance of a document. Subsequently, in a second experiment, we investigate reliable changes in the physiology due to reading texts with affective content. Together, the experiments comprehensively tackle physiological annotation for textual content.

founding factors in the physiological responses from emotional reactions of the users. We recorded the physiological responses of 40 participants while they read scientific documents, and asked them to provide relevance feedback. We investigated how perceived relevance was associated with peripheral physiology by analyzing the event-related physiological responses. Results show that electrodermal activity is different when users are reading relevant content than when they are reading irrelevant content.

Physiological annotation for affect. The second experiment was designed to study the physiological correlates of different affective states that resulted from reading relevant text items. The experiment was designed in the context of reading news articles. We recorded the physiological responses of 24 participants while they were allowed to freely browse the news from their favorite news website. In this more realistic setting users read only relevant information items, allowing physiological annotation for affective responses to relevant information items. We investigated how different affective responses were associated with peripheral physiology by analyzing the changes in physiological data while reading a news article. Results show that users present lower electrodermal activity when reading content eliciting negative emotional responses, than when reading emotionally neutral content.

The experiments on physiological text annotation show that annotating text items with physiological signals can help indicate the users' information needs and affective responses. The remainder of this paper is organized as follows: Section 2 overviews the background on implicit annotation of digital content through physiological signals. Next, Section 3 reports the experimental setup and results on physiological annotation for relevance, and Section 4 provides the experimental setup and results on physiological annotation for affect. In Section 5 we discuss the implications of the results and point out directions for future work. Finally, we conclude in Section 6.

2 Physiology for implicit annotation

Implicit annotation of content, also referred to as implicit tagging, has been studied in a wide range of research fields. Physiological signals are good candidates for implicit annotations. For instance, electrodermal activity (EDA) is related to the activation of the sympathetic nervous system and has largely been used as an indicator of emotional, psychological and emotional arousal (Andreassi, 2000). Other candidate signals include facial expressions, variations in the skin temperature, measures of neural activity and facial muscle activity.

Applications for affective annotation of content range from affective recommender systems to sentiment analysis and opinion mining. For instance, services such as the most read Italian daily newspaper (both on the internet and in paper) *Corriere.it* have implemented emotional annotation of their content through explicit user feedback. The service uses annotations to provide estimates of the general emotional response to a particular news piece. Another application for emotional annotation is found in the field of affective recommender systems (Tkalcic et al., 2010). These systems base their recommendations on affective metadata, derived from affective annotations of the content. Soleymani and Pantic (2012) have extensively reviewed the state of the art of physiological and other context-based metrics to implicitly annotate information items.

2.1 Physiology associated with relevance judgments

When engaging in an information seeking task, users perceive as relevant items that adjust to their information needs. In information retrieval, situational relevance is manifested through how aligned are the results to the user search intents. In order to maximize the relevance of results, it is common that information retrieval systems use relevance feedback mechanisms. Annotating content with some measure of relevance might be especially useful to filter or personalize the content to users. In this context, gathering relevance feedback on a set of results allows better personalization of the next iteration of results.

One way to gather relevance feedback is through direct user interaction, also known as explicit relevance feedback (Koenemann and Belkin, 1996). However, this measure has proven to be insufficient due to the burden it places on the user (Kelly and Fu, 2006). Another way to obtain relevance feedback is through unobtrusively monitoring the user, known as implicit relevance feedback (Kelly and Teevan, 2003). Research in the field of implicit relevance feedback through monitoring the user's state with physiological recordings is still at an early stage, especially regarding text content.

Arapakis et al. (2010) studied facial expressions as extracted from video recordings alone, or in combination with peripheral physiological measures, including electrodermal activity and skin temperature (Arapakis et al., 2009a), in order to infer the relevance of information items (namely documents and videos). They found that including affective features into their models increased the relevance prediction accuracies of their models based on implicit relevance metrics.

Moshfeghi and Jose (2013) studied whether affective (i.e., facial expressions from video recordings) and physiological signals (including heart rate, skin temperature and neural activity), in combination with traditional implicit relevance feedback metrics (i.e., dwell time), lead to improvement of relevance predictions. They studied a video retrieval engine over several search intentions, and found out that combining physiological and affective signals with traditional implicit signals led to significant improvement in the predictions.

Eugster et al. (2014) recently studied whether the relevance of terms could be predicted from brain signals alone. They measured neural activity through electroencephalography (EEG) while users were shown keywords. While carrying out a fairly controlled study, they were able to show that it is possible to infer the relevance of terms from brain signals only, resulting in significant improvements in the predictions over the random baseline.

2.2 Physiology associated with affective judgments

The role of emotions in information sciences has been studied extensively (Lopatovska, 2009; Lopatovska and Arapakis, 2011). Affective and physiological signals have been used as additional channels to implicitly measure human emotion and map emotional states (Pantic and Rothkrantz, 2003; Ioannou et al., 2005; Conati and Maclaren, 2009; Koelstra et al., 2012).

Koelstra et al. (2010) used electroencephalography as well as peripheral physiology as means to implicitly annotate music videos with the user's emotional state. They carried out a classification experiment in order to implicitly annotate the videos with measures of arousal and valence, reporting classification accuracies up to 80% in some cases.

Soleymani et al. (2008) studied the relationship between the physiological responses of participants while watching movie scenes, and their felt emotions. They inferred emotions from electrodermal activity, blood pressure, respiration rate, and facial electromyography. They found significant correlations between physiological features and video scenes, as well as between audio-visual features and emotions.

Affective and physiological signals have been used in the field of recommender systems, in combination with traditional implicit relevance measures (Arapakis et al., 2009b). The approach is to use affective and physiological measures as direct input for the recommendations. That is, physiological and affective signals are used to provide implicit affective feedback, which is directly used by the system to generate the recommendations. These recommender systems are known as affective recommender systems (Tkalcic et al., 2010, 2011). In affective recommender systems, it is common that emotions are inferred through affective features prior to their use as contextual information for generating the recommendations (Gonzalez et al., 2007).

Tkalcic et al. (2010) used affective metadata (metadata related to users' affective and emotional states) in image recommender systems. They studied affective metadata in combination with generic metadata (such as image genre or average watching time) for improving image recommendation results. Overall, their results showed better performance of the image recommender system when including affective metadata compared to using generic metadata alone.

Tkalcic et al. (2011) defined a framework for affective recommender systems. The framework formalizes the use of emotions in recommender systems in a three-stage model. The first stage (*entry stage*) refers to the emotional state of the user prior to the consumption of an information item. The second stage (*consumption stage*) refers to the emotional state of the user as influenced by the information item being consumed. The last stage (*exit stage*) refers to the emotional state of the user after having consumed the information item. The authors discussed the implications of modeling affect and emotions

2.3 Contributions to the field

We report on two experiments that explore physiological text annotation to extract measures of relevance and affect. Our work differs from the most related work in the following points:

1. *We study the physiological annotation of text*. Related work mostly has studied physiological signals to implicitly annotate either image or video content. However, a large amount of information is consumed in text format. We study physiological signals to implicitly annotate text content.
2. *We use physiology alone*. Most of the related work that uses physiological signals to implicitly annotate content uses physiological measures in combination with other signals such as facial expressions or behavioral signals (e.g., dwell time). Whereas these are totally valid approaches, these signals can be faked and sometimes might fail to capture the real level of user engagement with the system. We therefore explore which implicit judgments could be potentially derived from physiological annotation.
3. *We provide a comprehensive approach for physiological text annotation*. We first present a controlled experiment to study physiological text annotation for relevance, which is a central property regarding the users' information needs. Then, we study physiological text annotation for affect, in a much more ecologic environment. This provides a comprehensive understanding of how physiological signals can be used to implicitly annotate text content by first studying if a text item is likely to be consumed (relevance) and then the affective responses that relevant text items elicit (see Figure 1). We show in detail that electrodermal activity when reading relevant text is different than when reading irrelevant text, and that reading texts that make the user feel angry or sad present lower levels of electrodermal activity as compared with text that do not elicit emotional reactions. These are promising results for implicit text annotation from physiology, holding implications for several fields, namely for personalized user modeling, recommender systems and information retrieval.

3 Experiment 1: Physiological annotation for relevance

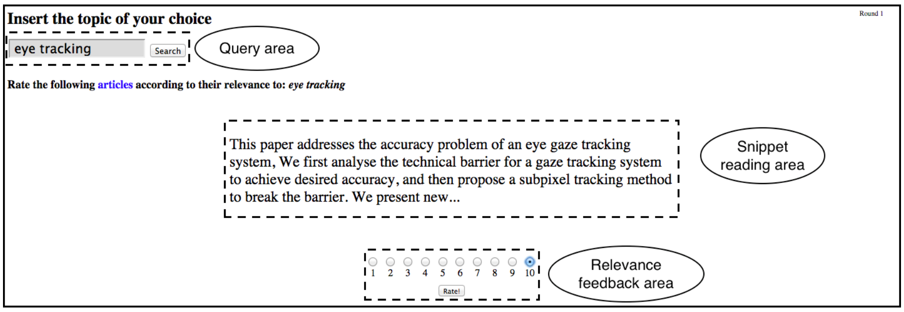
The first component that we addressed for physiological text annotation is relevance, as this is the most underlying property of the users' information needs, when engaged in information seeking tasks. The first experiment was therefore designed to investigate the physiological correlates of relevance in text reading tasks. The affective state of the user and the affective reactions to the information items being consumed can have a direct influence on the perception of relevance of the information items. For instance, the content of an article might elicit some affective response that in turn might influence the users relevance judgment on that article. The aim of the experiment was to minimize the role of affect in the relevance judgment process, in order to better study the physiological correlates of the relevance judgment. We therefore designed a relevance assessment task in the academic search context, as it was considered a fairly emotionally neutral context. This minimized the possibility of affect correlates conflating the relevance correlates, allowing to better capture the physiological responses to relevance manifestations, in terms of relevance feedback.

For the physiological recordings we selected two promising physiological signals that could be measured in a very non-intrusive manner. The first was electrodermal activity. This signal was especially well suited for our setup, as it has been used as an indicator of stimulus novelty, intensity, and emotional content, among others (Ravaja, 2004). Additionally, we measured corrugator supercilii activity (or CSA, activity over the brow muscle), as it has been associated with mental workload, fatigue, and compensatory mental effort, among others (Van Boxtel and Jessurun, 1993; Waterink and Van Boxtel, 1994; Veldhuizen et al., 2003).

We designed a relevance assessment experiment in which participants queried an information retrieval system indexing a scientific database and examined the set of results delivered. For each of the results, they were asked to judge the relevance to their query, while their physiological responses were being measured. The goal of the experiment was to find out whether the physiological annotations of content perceived as relevant reliably differed from the physiological annotations of content perceived as irrelevant.

3.1 Participants

Forty participants (six females) took part in the study. Participants ranged from 21 years old to 47 years old (M = 28.17) and were recruited after ensuring that they had previous experience in browsing scientific databases. Participants were students and staff from the University of Helsinki and Aalto University in Finland (37 participants were postgraduate, and three were undergraduate students). Only one of the participants reported being a native English speaker, and 17 different mother tongues were reported. Nevertheless, the overall English reading skills were self-reported as advanced. All of the participants used the mouse with their right hand, as they reported to be fully habituated to doing, even if six of them were left-handed. Participants reported themselves to be physically and mentally healthy.



Insert the topic of your choice

eye tracking

Rate the following articles according to their relevance to: *eye tracking*

This paper addresses the accuracy problem of an eye gaze tracking system, We first analyse the technical barrier for a gaze tracking system to achieve desired accuracy, and then propose a subpixel tracking method to break the barrier. We present new...

Fig. 2: Experimental task and user interface. The participant submits a query, and the system retrieves six abstract snippets of which three are relevant and three are irrelevant to the query. These are presented for the participant one at a time in a randomized order. The participant rates each result using a 1–10 scale. This is repeated for a total of six queries.

3.2 Apparatus

Physiological signals were recorded at a sample rate of 1000 Hz using a QuickAmp (BrainProducts GmbH., Germany) amplifier. CSA sensors were filled with SYNAPSE conductive electrode cream (Kustomer Kinetics Inc., USA) and placed on sites overlaying the left corrugator supercilii muscle regions as recommended by Fridlund and Cacioppo (1986). EDA electrodes were filled with TD-246 skin conductance electrode paste (Med Associates Inc., USA) and attached to the middle phalanges of the ring and little fingers of each participant's left hand (Dawson et al., 2007).

We built a custom search engine and user interface to have full control over the retrieval process, the presentation of results, and the content indexed by the retrieval system. The content items were from a scientific article database consisting of over 50,000,000 articles from the Web of Science prepared by Thomson Reuters, Inc. and from the Digital Libraries of the ACM, the IEEE, and Springer. We used a state-of-the-art unigram language modeling approach with Bayesian Dirichlet smoothing to rank the results (Zhai and Lafferty, 2004). The relevant results were retrieved directly based on the ranking provided by the ranking model. The irrelevant results were selected randomly with an additional Boolean constraint to exclude results that contained words from the participant's query. The first 40 words of the abstracts of the articles were used as result snippets, which we had found in pilot studies to be sufficient for participants to decide whether an article is relevant.

The results were presented in the Google Chrome browser. Custom JavaScript code was injected to record the exact time in milliseconds, in relation to the PC clock, when each event took place. To ascertain the synchrony between the browser timestamps and the physiological data, every second the experiment PC sent a synchronization pulse through the parallel port to the QuickAmp amplifier, and each of the browser events was synchronized to the closest pulse.

3.3 Procedure and experimental task

At the beginning of the session, the participants were briefed as to the procedure and purpose of the experiment, before signing informed consent. They were furthermore informed of their right to withdraw from the experiment at any time without any negative consequence. No training session was provided prior to the task, as the interaction with the system was particularly intuitive. After performing the task, the participants were asked to fill in an online survey regarding their background information, and their participation was compensated with two movie tickets.

The relevance assessment task was designed so that participants could perform queries of their interest using a real information retrieval system. The participants were presented with a search box and instructed to perform a query on a topic with which they were familiar, keeping in mind that they were browsing a scientific database (e.g., "eye tracking" as shown in Figure 2). Then, the search engine returned, in randomized order, six abstract snippets, one at a time. Of these six snippets, three were always actual search results (relevant to the participant's query) and the other three were randomly generated (irrelevant to the participant's query). The participants then read the abstract snippet and were asked to rate the relevance of the article on a scale from 1 to 10. Participants were instructed to provide the relevance feedback as soon as they made a decision on the relevance, without the need of reading the text until the end. On average, abstract snippets were shown for ca. 8 seconds. Then, the snippet was replaced with the next snippet until all 6 snippets corresponding to that query were rated and the participant was asked to perform a new query on a new topic. The experiment was completed after the participant rated 36 abstracts (6 topics x 6 abstract snippets).

3.4 Method and measures

After visual inspection, data from four out of the 40 participants was rejected due to loosened or misplaced sensors.

The physiological signals were exported using the BrainVision Recorder software and then filtered to reduce noise and artifacts (e.g., caused from the variation in the contact between the skin and the electrode). For EDA we used a low-pass filter with the cut-off at 5 Hz, and for CSA, a high-pass filter with the cut-off at 10 Hz. For both EDA and CSA signals, we executed the exact same feature extraction and analysis.

To annotate each abstract snippet, we extracted features from an 8-second window time-locked to the moment when the participant gave the explicit judgment. The window spanned from 2 seconds before to 6 seconds after the explicit rating. As we instructed the participants to assess the relevance as soon as they had made any decision on it, even though relevance-related processing may start earlier, the final decision on the relevance is expected to occur only shortly before the explicit rating. We therefore defined a time window that included two seconds before the explicit rating, which was considered to be sufficient to cover the time between the psychological decision and the physical event (Sternberg (1969) stimulus- response model); and six seconds following it, as skin conductance response (SCR) can take up to 6

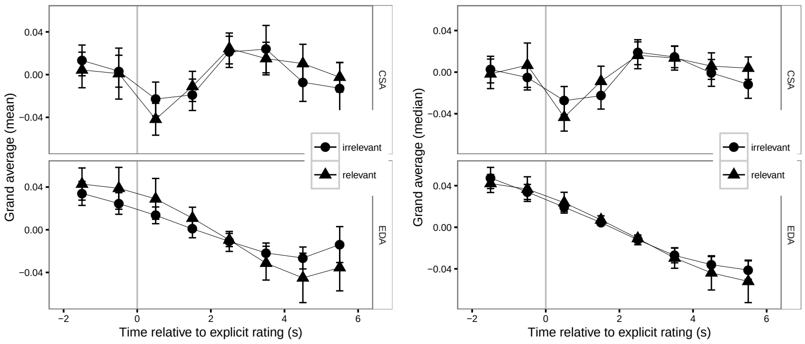


Fig. 3: Grand average based on mean values (left), and median values (right), with 95% confidence interval within the 8-second window of the electrodermal activity (EDA, bottom) and corrugator supercilii activity (CSA, top) signals averaged over participants and text snippets. Each point represents one of the features (i.e., averaged signal over 1 second). The vertical grey line at “0” indicates the explicit rating event.

or 7 seconds to reach its peak (Dawson et al., 2007). For every second within this window, we computed the average signal value (i.e., average down-sampling). This resulted in eight features characterizing EDA, and eight features characterizing CSA, for each abstract snippet.

The data for the participants were normalized so that the signals were centered on each participant’s mean level, and the features were divided by the standard deviation of the feature for each participant.

We compared the EDA and CSA responses of the participants for snippets that received different relevance feedback. The data were binarized to irrelevant and relevant categories. When the relevance judgment explicitly acquired from the participant was less than 4, it was categorized as an irrelevant judgment, and when the relevance judgment was higher than 7, it was categorized as a relevant judgment. Other snippets were not categorized, as the judgments were interpreted as ambiguous and the physiological responses associated to them were likely to be misleading.

Within each participant, we aggregated features from relevant and irrelevant snippets with both the arithmetic mean and the more robust median, and computed repeated-measures analysis of variance (ANOVA). The analyses were carried out using the General Linear Model (GLM) with repeated measures procedure in SPSS (version 21) with two-level factor “relevance” (relevant vs. irrelevant) and eight-level factor “time” (eight features corresponding to the eight seconds of the window). The results of the computed ANOVAs were corrected using the Greenhouse-Geisser correction on the degrees of freedom, as Mauchly’s Test indicated that the assumption of sphericity had been violated.

Table 1: ANOVA results corrected using the Greenhouse-Geisser correction on the degrees of freedom. Bold entries indicate statistical significance at the levels of $\alpha = .05$ (*), and $\alpha = .001$ (**).

(a) Electrodermal activity

	mean (EDA)			median (EDA)		
Source	df	MS	F	df	MS	F
Relevance	(1 , 35)	.00	.00	(1 , 35)	.00	**6.19** *
Time	(1.46 , 51.25)	.28	**20.56** **	(1.13 , 39.40)	.54	**56.44** **
Rel. * Time	(1.36 , 47.57)	.02	2.47	(1.26 , 44.05)	.00	1.32

(b) Corrugator supercilii activity

	mean (CSA)			median (CSA)		
Source	df	MS	F	df	MS	F
Relevance	(1 , 35)	.00	.00	(1 , 35)	.00	2.11
Time	(4.26 , 148.99)	.04	**5.97** **	(3.83 , 134.04)	.04	**8.99** **
Rel. * Time	(3.59 , 125.82)	.01	1.10	(4.35 , 152.38)	.00	1.57

3.5 Results

Figure 3 shows the aggregated features over participants for relevant and irrelevant articles using both the arithmetic mean and the median, with 95% confidence intervals, for the two signals. Table 1 presents the results of each of the four ANOVA tests.

ANOVA based on the mean EDA values showed a significant main effect of time, $F(1.46, 51.25) = 20.56, p < .001$. The ANOVA based on median EDA values showed a significant main effect of relevance, $F(1, 35) = 6.19, p = .018$, as well as time, $F(1.13, 39.40) = 56.44, p < .001$. The ANOVA based on the mean CSA values showed a significant main effect of time, $F(4.26, 148.99) = 5.97, p < .001$. Finally, The ANOVA based on the median CSA values showed a significant main effect of time, $F(3.83, 134.04) = 8.99, p < .001$.

The results of the ANOVAs can be summarized as follows: For both EDA and CSA, the main effect of time was found, which indicates that the physiological signal changes reliably following the relevance judgment. However, the direction of the judgment was only significant for EDA, as indicated by the significant main effect of the relevance. This means that decision-related physiological changes in CSA are either not related to perceived relevance, too weakly related to become visible, not stable enough across time to cause an interactive effect, or not the same between participants.

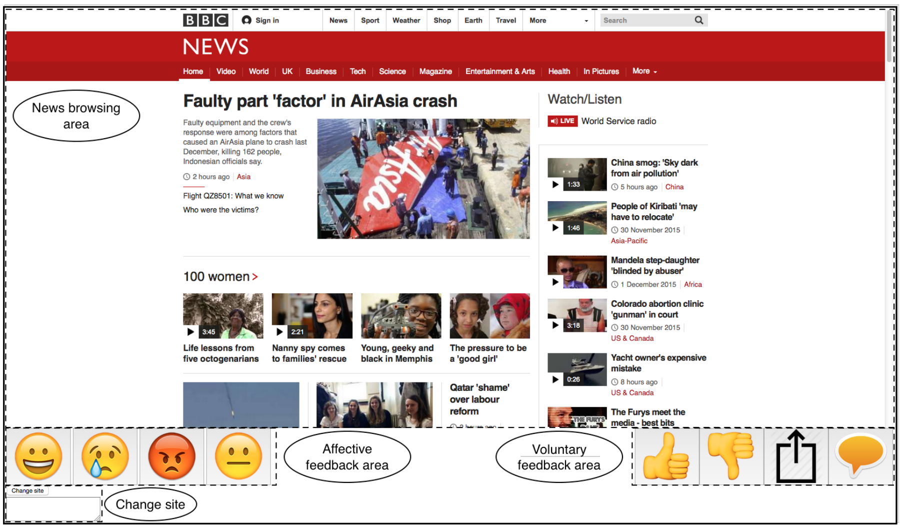
4 Experiment 2: Physiological annotation for affect

The experiment reported in Section 3 was designed as a fairly controlled experiment in which, in addition to the relevant results, we introduced irrelevant results to the users. This allowed to study the physiological correlates of relevance in a balanced setup. In the experiment we asked the participants to provide feedback as soon as they were able to judge the relevance of the text content. This allowed analyzing the specific physiological signals time-locked to the cognitive event of making a decision on the relevance of the item, as indicated by the explicit relevance feedback event.

After deciding that a text is relevant, one might start reading the relevant text item, which in turn might elicit several emotional responses (see Figure 1). We designed a second experiment to study physiological annotation for such contexts. This presents several challenges respect to the Experiment 1. First, it is very unlikely that the relevant information consumed by the participant will be balanced on the affective level (e.g., the participant most probably will consume a different amount of happy and sad articles). Therefore, one major challenge of the present experiment is to study physiological text annotation under unbalanced datasets. Second, affective responses might be elicited at different points during the reading process. Therefore, it is hard to analyze the physiological responses time-locked to a specific cognitive event as done in Experiment 1. In Experiment 2 we use a different approach, in which we include for the analysis the physiological data of the whole time-series from the beginning of the reading to the end of the reading, rather than just concentrating on a time-locked, event-specific time-window. This presents an additional level of complexity as, while measuring the effect of multiple event-related physiological responses to several events that are related to the text being processed, one might also be measuring the physiological responses to completely unrelated events (e.g., an external distraction).

The second experiment was therefore designed to study the physiological correlates of the affective responses of users when consuming relevant text items, as manifested in a news reading task (see Figure 1). We selected news reading, as it is a type of text content that has intrinsically associated emotional content, while keeping the user engaged and interested in the content. Participants could freely read the news of their choice from their favorite online news sources. This ensured that all the text content consumed by the users was relevant to them. It is important to note that we had no prior knowledge on the affective content of the articles being consumed. For each of the articles, participants were instructed to provide affective feedback according to how they felt after reading the article. The goal of the experiment was to find out whether several affective categories could be reliably associated with distinct physiological annotations. In order to study how physiological responses were associated with the participants perceived affective content of the news articles, we compared physiological responses to articles where the participant felt happy, sad, or angry against articles where the participant felt neutral.

In addition, we collected voluntary feedback on categories usually found on news sites: “thumbs up”, “thumbs down”, “comment” and “share”, and studied whether this type of feedback was correlated with the physiological signals. We also collected personality traits of the participants through the BIS/BAS questionnaire, in order to study the correlation between their personality and the type of feedback given to news articles. That is, we wanted to reveal whether people with certain personality profiles were more likely to give certain types of affective feedback, given that psychological and personality traits have been shown to influence the way people access and share content (Celli et al., 2015).



Faulty part 'factor' in AirAsia crash

Faulty equipment and the crew's response were among factors that caused an AirAsia plane to crash last December, killing 162 people, Indonesian officials say.

2 hours ago | Asia

Flight QZ8501: What we know

Who were the victims?

100 women >

3:45 Life lessons from five octogenarians

2:21 Nanny spy comes to families' rescue

Young, geeky and black in Memphis

The pressure to be a 'good girl'

Qatar 'shame' over labour reform

Watch/Listen

LIVE World Service radio

1:33 China smog: 'Sky dark from air pollution' 5 hours ago | China

1:46 People of Kiribati 'may have to relocate' 30 November 2015 | Asia-Pacific

Mandela step-daughter 'blinded by abuser' 1 December 2015 | Africa

3:18 Colorado abortion clinic 'gunman' in court 30 November 2015 | US & Canada

0:26 Yacht owner's expensive mistake 8 hours ago | US & Canada

The Furys meet the media - best bits

Fig. 4: Experimental task and user interface. The participant selects a news portal of her choice, and browses the news freely. After reading a specific news article, the participant clicks on one of the affective feedback icons (from left to right: "happy", "sad", "angry", and "neutral"). The participant is allowed to provide voluntary feedback, as well as to change a news portal at any point, by entering a new URL in the text box designated for it. Participants read news articles for 45-60 minutes.

Electrodermal activity was used as the physiological signal, as it has been proved to be indicative of arousal and stimulus novelty (Dawson et al., 2007; Boucsein, 2012). In order to further minimize the intrusiveness of the recording, and given the less binding findings for CSA reported in Experiment 1 (see Section 3), in the present experiment we solely relied on EDA.

4.1 Participants

Twenty-four participants (five females) took part in the study, two of which participated also in Experiment 1. Participants ranged from 23 to 36 years old (M = 29.7). Three participants were postdoctoral researchers, and the rest were students (18 postgraduate, and three undergraduate) from the University of Helsinki and Aalto University in Finland. Nine of the participants read news in their native language only, five read the news both in their native and foreign language, and ten only in a foreign language. In total, 15 different mother tongues were reported. Overall, participants reported high engagement with the content they were reading (M=4.08, on five-level Likert scale), and to not feel intruded by being asked to provide feedback (M = 1.95, on a five-level Likert scale). Two of the participants were left handed, even though only one of them used the computer mouse with the left hand. Participants reported themselves to be physically and mentally healthy.

4.2 Apparatus

A ProComp Infinity (Thought Technology Ltd., Canada) amplifier recorded EDA at a sample rate of 2048 Hz. Sensors were placed on the middle phalanges of the index and middle fingers of each participant's free hand (the one they did not use for the computer mouse) (Dawson et al., 2007). Stimulus were presented using the Google Chrome Browser.

The news was presented inside an overlay that allowed for affective feedback (see Figure 4). The overlay was implemented using a proxy Web server (Apache 2.4.10) that provided a HTML-page containing an iFrame tag that pointed to the external news site, as well as the buttons necessary for the user feedback. Whenever the user gave feedback using the overlay, the current time on the local machine was sent to the proxy server using an AJAX request. A similar mechanism was used to log the timestamps when the users changed Web pages. These timestamps used the same clock as the software running the physiological amplifier recording software, thus synchronizing the affective feedback with the physiological recording.

4.3 Procedure and experimental task

At the beginning of the session, participants were welcomed and briefed as to the procedure and purpose of the experiment before signing the informed consent form. Participants were instructed about the duration of the experiment and reminded that they could withdraw from the experiment at any point in time, without facing negative consequences. Before starting the recording, the experimenter briefly demonstrated the functioning of the experiment ensuring that the participant understood the experimental task. Participants were instructed to browse news sites of their interest (typically between 1 and 3) for a minimum of 45 minutes, and up to 60 minutes (M = 47.7, SD= 4.2). For each article read, they were asked to click one and only one of the affective icons, according to the way they felt reading the article. Participants could also provide voluntary feedback, clicking on as many icons as they pleased, including none (e.g., "thumbs up"; see Figure 4). Each participant read ca. 20 articles (M = 22.1, SD= 11.4), and spent ca. one and a half minutes per article (M = 1.4, SD = 1.3). While participants were allowed to read the text and look at the images on the articles, they were explicitly instructed to avoid playing videos. After the reading session, the participants were required to fill out a profiling questionnaire (see Section 4.4.2). Finally, they filled in an online survey regarding their background information, and their participation was compensated with one movie ticket.

4.4 Method and measures

4.4.1 Physiological analysis

The data from 19 participants were considered for the analysis after rejecting four participants that presented completely flat electrodermal recordings, possibly due to sensors misplacement or that they were "EDA non-responders" (Venables and Mitchell,

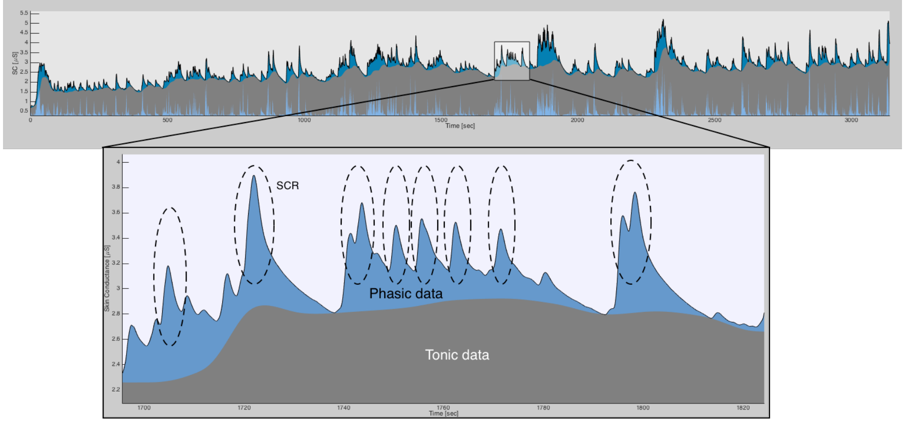


Fig. 5: Decomposed electrodermal activity (EDA) of one of the participants while reading one article. Grey and blue areas represent tonic and phasic activity respectively. Dashed ellipses represent skin conductance responses (SCRs). EDA was decomposed through discrete decomposition analysis using Ledalab software (Benedek and Kaernbach, 2010).

1996). Also, one participant, who had an abnormally low amount of feedback given (rating only five news articles) and spent drastically more time on each article, was removed from the analysis as an outlier.

EDA was exported using the BioGraph INFINTI software down-sampled to 32Hz. The signal was then separated into *phasic*, and *tonic* components, through continuous deconvolution analysis as provided by the MATLAB-based Ledalab software (Benedek and Kaernbach, 2010). The *phasic* component refers to electrodermal responses to stimuli and is characterized with spikes in the signal (i.e., skin conductance response; SCRs), whereas the *tonic* activity represents the skin conductance level (Boucsein, 2012).

To annotate each article read, we computed four features from the EDA signal: a) The sum of the tonic activity per second, computed as the sum of tonic data within an abstract (centered with the mean tonic data of the first 10 seconds spent reading the article) divided by the time spent reading the article (*sumTonic*), b) the sum of the phasic activity per second, computed as the sum of the phasic data within an article, divided by the time spent reading the article (*sumPhasic*), c) the average amplitude of the SCRs, computed as the sum of the SCRs' amplitudes within an article divided by the number of SCRs in the article (*meanAmpSCR*), and d) the number of SCRs per second, computed as the number of SCRs divided by the time spent reading the article (*nSCR*).

The data for the participants were normalized so that the signals were centered on each participant's mean level, and the features were divided by the standard deviation of the feature for each participant.

We compared the EDA responses of the participants for news articles that received different affective feedback. The 19 participants considered for the analysis read a total 425 articles, of which 133 were rated as "happy", 44 as "sad", 68 as "angry", and 180 as "neutral". Additionally, the 19 participants also provided voluntary feedback on 132 articles. Specifically, 74 articles were "liked", 20 were "disliked", 64 would have been "shared", and 43 would have been "commented".

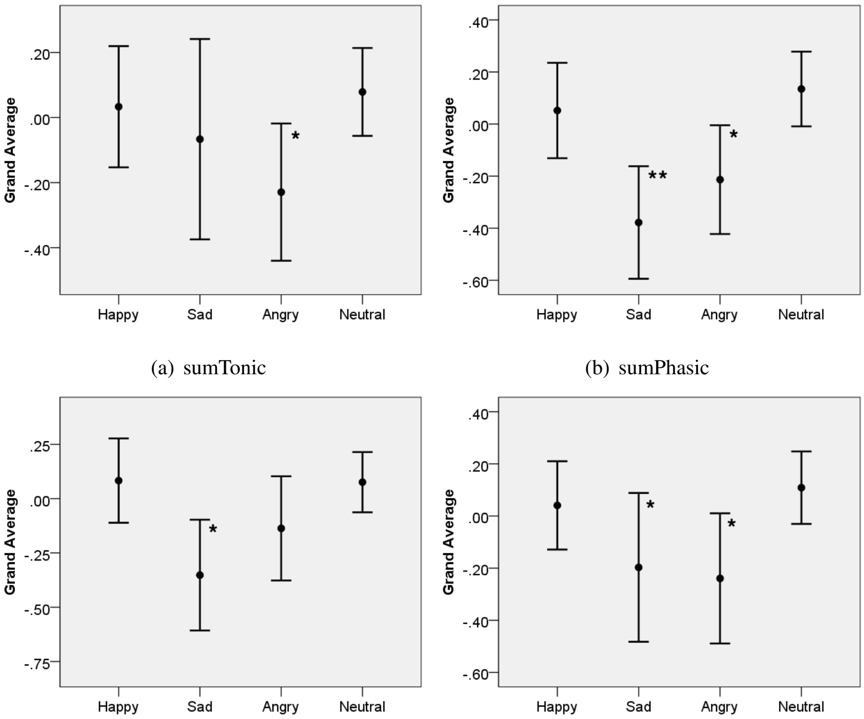
Instead of the repeated measures ANOVA that was used in Experiment 1 (see Section 3.4), we used the Linear Mixed Models (LMM) analysis owing to the unbalanced nature of the data. That is, given that the participants had the freedom to select the articles to read and to spend any time on each article, the participants had varying numbers of rated articles. Mixed models have been shown to handle missing observations and unbalanced designs more efficiently, resulting in more reliable conclusions (Bagiella et al., 2000). They can also lead to more efficient estimates and more powerful tests as they involve a model for the error variance.

The statistical analyses were carried out using the Linear Mixed Models procedure in SPSS (version 21) with restricted maximum likelihood estimation and a first-order autoregressive covariance structure for the residuals. Participant ID was specified as the subject variable. Article number, an integer variable that indicates the order at which each user read their articles, was specified as the repeated variable. One at a time each of EDA features were selected as the dependent variable and the model included a fixed effect for the affective feedback category as the independent variable. To analyze the affective feedback, we used the "neutral" affective feedback as the control group, and compared it using planned contrasts against the three other categories: "happy", "sad", and "angry".

To analyze the voluntary feedback a new binary variable was created for each of the four types of voluntary feedback, "thumbs up", "thumbs down", "share" and "comment", to indicate whether an article had received this type of feedback. Also, a variable was created to indicate whether an article had received any of the four types of voluntary feedback. Then the same Linear Mixed Models procedure was run with each of these indicator variables as the independent variable. Thus, we compared all the articles that had received "thumbs up" to those articles that had not received "thumbs up". Similar analysis was done for "thumbs down", "share" and "comment". Finally, we compared all the articles that had received any of the four types of voluntary feedback to those articles that had not.

4.4.2 Profiling questionnaire

To profile the users, we used the BIS/BAS questionnaire which is based on the assumption that two general motivational systems underlie behavior. One system is responsible for aversive motives while the other regulates appetitive motivations (Gray and McNaughton, 2003). The aversive part of the model is called behavioral inhibition system (BIS) and is sensitive to signals of punishment, non-reward, and novelty: it inhibits behavior that might lead to negative or painful outcomes. Therefore, BIS activation causes inhibition of movement toward goals: "In terms of individual differences in personality, greater BIS sensitivity should be reflected in greater proneness to anxiety, provided the person is exposed to the proper situational cues." (Carver and White, 1994). In more colloquial language, people who are inhibited in this sense are more sensitive to possible negative environmental stimuli which in turn can lead to



(a) sumTonic

(b) sumPhasic

(c) meanAmpSCR

(d) nSCR

Fig. 6: Grand average with 95% confidence intervals for each of the four EDA features and affective feedback categories, with signals averaged over participants and articles. Asterisks indicate statistical difference respect to the "neutral" category, at the levels of $\alpha = .05$ (*), and $\alpha = .001$ (**).

avoidance behavior. Similarly, any novel stimulus is a potential risk that might lead to a painful outcome, thereby causing activation of the inhibition system.

The part of the model responsible for appetitive motivation is called the behavioral approach system and is said to be sensitive to signals of reward, non-punishment and escape from punishment. This system motivates the person to move toward goals, and is also responsible for the positive feelings, such as hope, elation and happiness. "In terms of individual differences in personality, greater BAS sensitivity should be reflected in greater proneness to engage in goal-directed efforts and to experience positive feelings when the person is exposed to cues of impending reward." The questionnaire measures BIS with a single index and divides the BAS part of the system into three categories: Drive, Reward Responsiveness and Fun seeking.

The data from 23 participants was considered for the analysis, as the participant who had given feedback to only 5 articles was removed from the analysis as an outlier. In the experiment we studied whether the BIS/BAS scales correlate with the affective feedback given to articles. As the participants rated varying numbers of articles we

Table 2: Results of the Linear Mixed Models (LMM) analysis for each of the four features, comparing each of the affective categories against the "neutral" control category using planned contrasts. Bold entries indicate statistical significance at the levels of $\alpha = .05$ (*), and $\alpha = .001$ (**).

	Happy vs. Neutral		Sad vs. Neutral		Angry vs. Neutral	
Feature	t	df	t	df	t	df
sumTonic	.49	401.40	.73	390.15	**2.16***	399.32
sumPhasic	.84	420.89	**3.30****	419.77	**2.43** *	420.96
meanAmpSCR	.05	392.04	**2.50***	392.76	1.38	392.56
nScr	.65	421.00	**2.00***	417.77	**2.48***	420.67

calculated percentages of each category for each participant, for example a participant might have rated 30% of all their articles as "sad". We then calculated the Spearman correlation between these percentages and the BIS/BAS scores of the participants.

4.5 Results

4.5.1 Affective feedback

Figure 6 shows the aggregated features over participants for each of the affective feedback categories, using the arithmetic mean, with 95% confidence intervals. Table 2 details the results of the Linear Mixed Models analysis.

The results for the affective feedback can be summed up as follows: for all of the EDA features, there was no statistically significant difference between the "happy" and "neutral" categories. However, each EDA feature showed a significant decrease for "sad" or "angry" or both. Thus the rather interesting pattern emerges: news articles rated as eliciting negative emotions seemed to generate less arousal (see Figure 6). In detail, for the tonic feature the difference between "angry" and "neutral" was statistically significant t(399.32) = 2.16, p = .032. The "sad" and "happy" categories did not differ from the control "neutral" category. For the sum of phasic activity feature the difference between "sad" and "neutral" was statistically significant t(419.77) = 3.30, p = .001. Also, the difference between "angry" and "neutral" was statistically significant t(420,96) = 2.43, p = .016. However, the difference between "happy" and "neutral" was not significant. For the mean amplitude feature the difference between "sad" and "neutral" was statistically significant t(392.76) = 2.50, p = .013. The value for "happy" was close to the control level and "angry" was somewhat below control, but the difference was not statistically significant. Finally, for the number of SCRs feature the difference between "sad" and "neutral" was statistically significant t(418.76) = 2.00, p = .047. Also, the difference between "angry" and "neutral" was statistically significant t(420.67) = 2.48, p = .013. The value for "sad" and "angry" were below the control "neutral". However, the difference between "happy" and "neutral" was not significant.

Table 3: Results of the correlation analysis, between the number of articles rated for each category and the BIS/BAS scores of the participants. Spearman's correlation coefficient is reported (r_s). Bold entries indicate statistical significance at the level of $\alpha = .05$ (*).

		BAS		
	BIS	Drive	Reward Responsiveness	Fun Seeking
Happy	**-.51** *	**.47** *	.16	.36
Sad	.24	-.18	.02	-.17
Angry	.34	.06	.31	.13
Neutral	-.09	-.16	**-.42** *	-.11

4.5.2 Voluntary feedback

The analysis did not yield statistically significant results for any of the voluntary feedback categories: "thumbs up", "thumbs down", "share" or "comment".

4.5.3 BIS/BAS questionnaire

Table 3 presents the results of each of the four correlation tests.

The results for the BIS/BAS questionnaire can be summed up as follows: The BIS part of the questionnaire that measures inhibition and aversion correlated negatively with the "happy" feedback, while the BAS part that measures appetitive motivations correlated positively with the "happy" feedback and negatively with the "neutral" feedback. In detail, The BIS correlated negatively with "happy" feedback (r_s = -.51, p = .013) indicating that inhibited participants rated articles as being "happy" less often. The BAS Drive correlated positively with the "happy" feedback (r_s = .47, p = .025), indicating that people driven to pursue positive things rated articles as "happy" more often (BAS Drive is measured with questions such as "I go out of my way to get things I want."). The BAS Reward Responsiveness was negatively correlated with the "neutral" category (r_s = -.42, p = .043) indicating that people who respond more strongly to positive environmental cues were less likely to rate the news article as "neutral". BAS Reward Responsiveness was measured with questions such as "When I get something I want, I feel excited and energized." Finally, as might be expected, BAS Fun Seeking category was positively correlated with the "happy" feedback, but this result failed to reach statistical significance (r_s = .36, p = .096). The BIS/BAS scales did not correlate with any of the voluntary feedback categories ("thumbs up", "thumbs down", "share", "comment"), and therefore are not included in Table 3.

5 Discussion

Overall, results show promising findings for physiological text annotation. The results from Experiment 1 indicate that the electrodermal activity when reading relevant text items is different than when reading irrelevant text items. On the other side, the results from Experiment 2 indicate, that "all arousal is good arousal", as both of the negative feedback categories elicited lower arousal than the control "neutral" category, and the positive "happy" category did not show difference with the "neutral" category.

5.1 Implications

The goal of this paper is to present evidence that implicit annotation of text content through physiological signals can reveal insights on the user's information needs and affective responses. The results point toward two direct implications for implicit text annotation with physiological signals. The first one is that annotating the textual content with the amount of electrodermal activity would help determine whether the document was relevant to the user. The second one is that annotating the textual content with measures of electrodermal activity could help determine whether the article has negative emotional content.

The advantages of implicit text annotation spans across several fields. Two of the most closely related fields are information retrieval and affective recommender systems. Accordingly, we discuss the implications of the results specifically in these two fields.

Implications for information retrieval. The results of Experiment 1 have direct implications for information retrieval. The results indicate that electrodermal responses to relevant and irrelevant articles are different. Accordingly, using electrodermal activity alone or in combination with other implicit feedback measures to annotate text results according to their relevance to the user should be, in principle, possible. We recognize these as preliminary but promising results for the field of information retrieval.

Implications for affective recommender systems. The findings for physiological text annotation have direct implications for affective recommender systems. These systems rely on affective metadata to generate their predictions. The results are highly encouraging in the direction of using physiological signals (namely electrodermal activity) as affective metadata for textual content. According to the framework for affective recommender systems developed by Tkalcic et al. (2011) – briefly overviewed in Section 2 – the results directly relate to the *consumption stage* of an affective recommender system, as we have been measuring the affective responses caused by the consumption of an information item. In addition, it should be studied how the results could generalize to other stages of the information consumption process, generating affective metadata for the *entry* and *exit* stages as well. That is, physiological measures could be used to generate affective metadata both to gain knowledge on the prior affective state of the user when starting to consume and information item (i.e., *entry stage*), and after having consumed it (i.e., *exit stage*).

5.2 Limitations and future work

The results show that electrodermal activity can be used for implicit text annotation, however, several important challenges remain for future work. Below we list a series of limitations of our approach, and possible directions for future work.

Relevance vs. affect. The experiments presented in this paper were specifically designed to minimize interaction effects between relevance and affect. For instance, Experiment 1 was designed as a relevance assessment task using scientific articles. In this way the user was concentrated in providing relevance judgment to the scientific articles, minimizing confounding factors in the physiology coming from external emotional responses. However, one cannot completely separate relevance from affect, as affect is an intrinsic part of the relevance judgment process. Furthermore, when users are engaged in more realistic information seeking tasks (rather than in simplified relevance assessment tasks), judging the relevance of information items becomes an intermediate step towards a greater goal (e.g., learning about a topic). In these more complex and exploratory information seeking tasks, affect is likely to have a much greater interaction with the relevance judgment process, as factors such as the overall search experience, system responsiveness or difficulty of the task might influence the user's affective state and, in turn, the way in which the information items are perceived. Such affective component of the relevance judgment process has been defined in information science as *affective relevance* (Saracevic, 1996). For instance, a user might feel annoyed by a system that does not match her expectations, in turn influencing the way she perceives the results of the system. Therefore, the emotional and affective state of the users has an effect on their perception of relevance of information items (Cosijn and Ingwersen, 2000), and physiology has been pointed out as a way to study this component of relevance (Barral and Jacucci, 2014). Future work research can extend results assessing even more realistic settings that consider more realistic information seeking scenarios where relevance and affect might be more intertwined, further helping to bring the notion of *affective relevance* from information science theory to real information systems.

All arousal is good arousal. The results of Experiment 2 seem to indicate that "all arousal is good arousal" as the negatively rated articles generated less arousal than the neutral category. To be more precise, as articles labeled as happy did not generate significantly more arousal than articles labeled as neutral, it can be reasoned that arousal indicates lack of negative affect. Moreover, it can be argued that some emotional states were not captured, as not all possible states were available for feedback. Therefore, there is the possibility that the control "neutral" state included some states that are not anger, sadness or happiness, but not really neutral either. Thereafter, future work could benefit from a more comprehensive affective feedback collection.

The role of valence. While electrodermal activity has several positive qualities such as non-intrusiveness, it is also a reasonably simple signal. For instance, recording corrugator supercilii activity or other facial muscles activity in Experiment 2 could have brought very valuable input for the physiological text annotation framework, as facial muscles are well known to be correlated with emotional valence (Cacioppo et al., 1986). This would allow to better cover the affective states on the valence-arousal dimension (Posner et al., 2005). Naturally, one should take into account the trade-off between the intrusiveness of the recording and the affective space that is being covered, and in the present work we primed the non-intrusiveness. However, exploring the combination of the results presented for electrodermal activity with other physiological signals would most likely enhance the potential of physiological text annotation, allowing to better map the emotional states through the valence-arousal space.

Additionally, the analysis of the voluntary feedback in Experiment 2 did not yield to significant results. The aim of the analysis was to capture the physiological responses to more complex emotional constructs that would lead the users to like, dislike, share, or comment a specific article. Recording a greater amount of physiological signals would increase the chances of capturing such more complex emotional constructs.

Personality traits. Initial results from the personality traits questionnaires indicate that it is possible to predict how different users provide affective feedback based on their personality profile (in this case the BIS/BAS questionnaire). It would be beneficial to study if this could be combined with the physiological results in more comprehensive way. Also, as in the experiment the participants were free to select the articles to read, it might be that participants with certain type of personality profile selected different types of articles to read, and not only that their personalities caused them to score the articles differently. Future work could study user models that take into account both the correlation between the affective feedback and the users' physiological responses, and how different types of personalities provide affective different feedback.

The class unbalance. One major challenge for the application of physiological text annotation is to study the accuracy of the annotation, in setups where the distribution of classes in the dataset might most probably be of an unbalanced nature. That is in fact the kind of dataset we gathered in Experiment 2, the results of which indicate that the affective categories that could possibly be annotated correctly using electrodermal activity are negative versus non-negative. However, the data gathered is heavily unbalanced, leaning towards non-negative contents. This might be problematic when addressing classification scenarios to assess the accuracy of the annotations, and is a valuable insight for future work studying classification setups for physiological annotation, as the distribution of classes is an important aspect to consider when implementing classification models. Furthermore, one might take into consideration collecting data using a predefined dataset, where the distribution of classes is known beforehand, in order to gather a more balanced and complete dataset where the accuracy of the annotation could be studied under more controlled experimental conditions.

Applications of physiological text annotation. Even though our results show that implicit annotation of text content is feasible using physiology alone, direct applications of physiological text annotation are still unclear. In fact, in the present submission we do not address the annotated physiological features for end-user applications. The intrusiveness of the recordings, the unbalance of the data, or the acceptance of the technology are aspects that need further study in order to implement physiological text annotation in end-user applications. Applications range from corpus annotation to personalized user modeling. The former would require to study measures of agreement in order to validate the annotations in terms of how generalizable are the physiological annotations. The later would require to study the extent in which the annotations are valid; that is, in a scenario where the information needs are dynamic, the perception of relevance towards an information item might change across stages of the information seeking process. Therefore, while the present paper presents the fundamentals for physiological text annotation, the accuracy and usefulness of the annotation in practical applications remains a subject for further experimentation.

Further, wearable technology that is able to measure electrodermal activity and other physiological signals in a non-intrusive manner is readily available (e.g.,Microsoft Band, `www.microsoft.com/microsoft-band`; Moodmetric ring, `www.moodmetric.com`). Future work should also address the measurement and processing of physiological signals through such devices, studying how to handle physiological annotations considering mobility or interruptions, when users are consuming textual information on the move. We believe that the present paper provides a useful starting point for developers who are planning to utilize the data generated by these devices in commercial applications.

6 Conclusion

We presented a conceptual framework for physiological text annotation, reporting on two experiments that studied how associating physiological responses to text content can indicate user's information needs and affective states (see Figure 1). We regard these experiments as the first natural steps for generating affective metadata of text items based on physiological measures only.

The first experiment studied the physiological correlates of relevance in information retrieval. Electrodermal activity and corrugator supercilii activity were recorded, and participants provided relevant feedback on the results. Results showed reliable differences in the electrodermal activity for relevant and irrelevant articles.

The second experiment studied the physiological correlates of affective states in text reading. Electrodermal activity was recorded while participants read news articles and provided feedback according to their affective state. Results showed lower electrodermal activity while reading negative articles (marked as sad or angry), as compared to emotionally neutral articles.

These results provide evidence that implicit annotation of text content using physiological signals is possible. The applications range from affective recommender systems, to implicit relevance feedback in information retrieval and human-centered implicit tagging.

References

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, June 2005. ISSN 1041-4347. doi: 10.1109/TKDE.2005.99.

John L. Andreassi. *Psychophysiology: Human behavior & physiological response*. Psychology Press, 2000.

Ioannis Arapakis, Ioannis Konstas, and Joemon M. Jose. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 461–470, New York, NY, USA, 2009a. ACM. ISBN 978-1-60558-608-3. doi: 10.1145/1631272.1631336. URL `http://doi.acm.org/10.1145/1631272.1631336`.

Ioannis Arapakis, Yashar Moshfeghi, Hideo Joho, Reede Ren, David Hannah, and Joemon M. Jose. Enriching user profiling with affective features for the improvement of a multimodal recommender system. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 29:1–29:8, New York, NY, USA, 2009b. ACM. ISBN 978-1-60558-480-5. doi: 10.1145/1646396.1646433. URL `http://doi.acm.org/10.1145/1646396.1646433`.

Ioannis Arapakis, Konstantinos Athanasakos, and Joemon M. Jose. A comparison of general vs personalised affective models for the prediction of topical relevance. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 371–378, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0153-4. doi: 10.1145/1835449.1835512. URL `http://doi.acm.org/10.1145/1835449.1835512`.

Emilia Bagiella, Richard P. Sloan, and Daniel F. Heitjan. Mixed-effects models in psychophysiology. *Psychophysiology*, 37:13–20, 1 2000. ISSN 1469-8986. URL `http://journals.cambridge.org/article_S0048577200980648`.

Oswald Barral and Giulio Jacucci. Applying physiological computing methods to study psychological, affective and motivational relevance. In Giulio Jacucci, Luciano Gamberini, Jonathan Freeman, and Anna Spagnolli, editors, *Symbiotic Interaction*, volume 8820 of *Lecture Notes in Computer Science*, pages 35–46. Springer International Publishing, 2014. ISBN 978-3-319-13499-4. doi: 10.1007/978-3-319-13500-7_3. URL `http://dx.doi.org/10.1007/978-3-319-13500-7_3`.

Oswald Barral, Manuel J.A. Eugster, Tuukka Ruotsalo, Michiel M. Spapé, Ilkka Kosunen, Niklas Ravaja, Samuel Kaski, and Giulio Jacucci. Exploring peripheral physiology as a predictor of perceived relevance in information retrieval. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 389–399, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3306-1. doi: 10.1145/2678025.2701389. URL `http://doi.acm.org/10.1145/2678025.2701389`.

Mathias Benedek and Christian Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1):80 –

91, 2010. ISSN 0165-0270. doi: http://dx.doi.org/10.1016/j.jneumeth.2010.04.028. URL http://www.sciencedirect.com/science/article/pii/S0165027010002335.

Wolfram Boucsein. *Electrodermal Activity*. Springer US, Boston, MA, 2012. ISBN 9781461411253; 9781461411260. URL http://edepot.wur.nl/207828.

J. T. Cacioppo, R. E. Petty, M. E. Losch, and H. S. Kim. Electromyographic activity over facial muscle regions can differentiate the valence and intensity of affective reactions. *Journal of personality and social psychology*, 50(2):260–268, February 1986. ISSN 0022-3514. URL http://view.ncbi.nlm.nih.gov/pubmed/3701577.

Charles S. Carver and Teri L. White. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The bis/bas scales. *Journal of personality and social psychology*, 67(2):319–333, 1994. ISSN 0022-3514. doi: 10.1037/0022-3514.67.2.319. URL http://dx.doi.org/10.1037/0022-3514.67.2.319.

Fabio Celli, Arindam Ghosh, Firoj Alam, and Giuseppe Riccardi. In the mood for sharing contents: Emotions, personality and interaction styles in the diffusion of news. *Information Processing & Management*, pages – , 2015. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/j.ipm.2015.08.002. URL http://www.sciencedirect.com/science/article/pii/S030645731500103X.

Cristina Conati and Heather Maclaren. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19(3): 267–303, 2009.

Erica Cosijn and Peter Ingwersen. Dimensions of relevance. *Information Processing & Management*, 36(4):533 – 550, 2000. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/S0306-4573(99)00072-2.

James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. The youtube video recommendation system. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 293–296, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-906-0. doi: 10.1145/1864708.1864770. URL http://doi.acm.org/10.1145/1864708.1864770.

Michael E. Dawson, Anne M. Schell, Diane L. Filion, and Gary G. Berntson. The electrodermal system. In John T. Cacioppo, Louis G. Tassinary, and Gary Berntson, editors, *Handbook of Psychophysiology*, pages 157–181. Cambridge University Press, third edition, 2007. ISBN 9780511546396. URL http://dx.doi.org/10.1017/CBO9780511546396.007. Cambridge Books Online.

Manuel J.A. Eugster, Tuukka Ruotsalo, Michiel M. Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. Predicting term-relevance from brain signals. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 425–434, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2257-7. doi: 10.1145/2600428.2609594. URL http://doi.acm.org/10.1145/2600428.2609594.

Alan J. Fridlund and John T. Cacioppo. Guidelines for human electromyographic research. *Psychophysiology*, 23(5):567–589, 1986.

G. Gonzalez, J.L. de la Rosa, M. Montaner, and S. Delfin. Embedding emotional context in recommender systems. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, pages 845–852, April 2007. doi: 10.1109/ICDEW.2007.4401075.

Jeffrey Alan Gray and Neil McNaughton. *The neuropsychology of anxiety: An enquiry into the function of the septo-hippocampal system.* Number 33. Oxford university press, 2003.

Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22 (1):5–53, January 2004. ISSN 1046-8188. doi: 10.1145/963770.963772. URL `http://doi.acm.org/10.1145/963770.963772`.

Spiros V. Ioannou, Amaryllis T. Raouzaiou, Vasilis A. Tzouvaras, Theofilos P. Mailis, Kostas C. Karpouzis, and Stefanos D. Kollias. Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Networks*, 18(4): 423 – 435, 2005. ISSN 0893-6080. doi: http://dx.doi.org/10.1016/j.neunet.2005.03.004. URL `http://www.sciencedirect.com/science/article/pii/S0893608005000377`. Emotion and Brain.

Diane Kelly and Xin Fu. Elicitation of term relevance feedback: An investigation of term source and context. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 453–460, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148249. URL `http://doi.acm.org/10.1145/1148170.1148249`.

Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2):18–28, September 2003. ISSN 0163-5840. doi: 10.1145/959258.959260. URL `http://doi.acm.org/10.1145/959258.959260`.

S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis ;using physiological signals. *Affective Computing, IEEE Transactions on*, 3(1):18–31, Jan 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2011.15.

Sander Koelstra, Ashkan Yazdani, Mohammad Soleymani, Christian Mhl, Jong-Seok Lee, Anton Nijholt, Thierry Pun, Touradj Ebrahimi, and Ioannis Patras. Single trial classification of eeg and peripheral physiological signals for recognition of emotions induced by music videos. In Yiyu Yao, Ron Sun, Tomaso Poggio, Jiming Liu, Ning Zhong, and Jimmy Huang, editors, *Brain Informatics*, volume 6334 of *Lecture Notes in Computer Science*, pages 89–100. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15313-6. doi: 10.1007/978-3-642-15314-3_9. URL `http://dx.doi.org/10.1007/978-3-642-15314-3_9`.

Jürgen Koenemann and Nicholas J. Belkin. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, pages 205–212, New York, NY, USA, 1996. ACM. ISBN 0-89791-777-4. doi: 10.1145/238386.238487. URL `http://doi.acm.org/10.1145/238386.238487`.

Geoffrey Leech. Introducing corpus annotation. In Roger Garside, Geoffrey Leech, and Anthony Mcenery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Addison Wesley Longman, 1997.

Irene Lopatovska. Searching for good mood: Examining relationships between search task and mood. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–13, 2009. ISSN 1550-8390. doi: 10.1002/meet.2009.1450460222. URL `http://dx.doi.org/10.1002/meet.2009.1450460222`.

Irene Lopatovska and Ioannis Arapakis. Theories, methods and current research on emotions in library and information science, information retrieval and humancomputer interaction. *Information Processing & Management*, 47(4):575 – 592, 2011. ISSN 0306-4573. doi: http://dx.doi.org/10.1016/j.ipm.2010.09.001. URL `http://www.sciencedirect.com/science/article/pii/S0306457310000737`.

Yashar Moshfeghi and Joemon M. Jose. An effective implicit relevance feedback technique using affective, physiological and behavioural features. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 133–142, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2034-4. doi: 10.1145/2484028.2484074. URL `http://doi.acm.org/10.1145/2484028.2484074`.

M. Pantic and L.J.M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, Sept 2003. ISSN 0018-9219. doi: 10.1109/JPROC.2003.817122.

Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(03):715–734, 2005.

Niklas Ravaja. Contributions of psychophysiology to media research: Review and recommendations. *Media Psychology*, 6(2):193–235, 2004.

Tefko Saracevic. Relevance reconsidered. In *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)*, pages 201–218. ACM Press, 1996.

M. Soleymani and M. Pantic. Human-centered implicit tagging: Overview and perspectives. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pages 3304–3309, Oct 2012. doi: 10.1109/ICSMC.2012.6378301.

M. Soleymani, G. Chanel, J.J.M. Kierkels, and T. Pun. Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 228–235, Dec 2008. doi: 10.1109/ISM.2008.14.

Saul Sternberg. The discovery of processing stages: Extensions of donders' method. *Acta Psychologica*, 30(0):276 – 315, 1969. ISSN 0001-6918. doi: http://dx.doi.org/10.1016/0001-6918(69)90055-9. URL `http://www.sciencedirect.com/science/article/pii/0001691869900559`.

Marko Tkalcic, Urban Burnik, and Andrej Koir. Using affective parameters in a content-based recommender system for images. *User Modeling and User-Adapted Interaction*, 20(4):279–311, 2010. ISSN 0924-1868. doi: 10.1007/s11257-010-9079-z. URL `http://dx.doi.org/10.1007/`

s11257-010-9079-z.

Marko Tkalcic, A. Kosir, and Jurij Tasic. Affective recommender systems: the role of emotions in recommender systems. In *Proc. The RecSys 2011 Workshop on Human Decision Making in Recommender Systems*, pages 9–13. Citeseer, 2011.

Erin Treacy Solovey, Daniel Afergan, Evan M. Peck, Samuel W. Hincks, and Robert J.K. Jacob. Designing implicit interfaces for physiological computing: Guidelines and lessons learned using fnirs. *ACM Trans. Comput.-Hum. Interact.*, 21(6):35:1–35:27, January 2015. ISSN 1073-0516. doi: 10.1145/2687926. URL http://doi.acm.org/10.1145/2687926.

A. Van Boxtel and M. Jessurun. Amplitude and bilateral coherency of facial and jaw-elevator emg activity as an index of effort during a two-choice serial reaction task. *Psychophysiology*, 30(6):589–604, 1993.

I.J.T. Veldhuizen, A.W.K Gaillard, and J. De Vries. The influence of mental fatigue on facial emg activity during a simulated workday. *Biological Psychology*, 63(1): 59–78, 2003.

P.H. Venables and D.A. Mitchell. The effects of age, sex and time of testing on skin conductance activity. *Biological Psychology*, 43(2):87 – 101, 1996. ISSN 0301-0511. doi: http://dx.doi.org/10.1016/0301-0511(96)05183-6. URL http://www.sciencedirect.com/science/article/pii/0301051196051836.

W. Waterink and A. Van Boxtel. Facial and jaw-elevator emg activity in relation to changes in performance level during a sustained information processing task. *Biological Psychology*, 37(3):183–198, 1994.

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.