Emotional Text Tagging

Farida Ismail
German University in
Cairo, Egypt
farida.ismail@gmail.com

Ralf Biedert, Andreas Dengel
German Research Center
For Artificial Intelligence
firstname.lastname@dfki.de

Georg Buscher
Microsoft
One Microsoft Way
Redmond, WA 98052, USA
georg@gbuscher.com

ABSTRACT

We created and evaluated a system capable of observing the reader's emotions and tag the perused text. By using either a web camera or an Emotiv neuroheadset displayed emotions like happiness, interest, boredom and doubt can be recorded. At the same time an eye tracker analyzes the reader's progress. According to the reader's current reading position in the text and the displayed emotions, the text part is automatically tagged with the reader's emotional state. The reading-interface is able to facilitate the emotional information in realtime, the user can also access the recorded eye tracking sessions and perceived emotions later on. We evaluated the system's ability to accurately tag emotions, conclude that joy is detected best, boredom is barely recognizable, and highlight some key issues we encountered.

Author Keywords
Eye Tracking, Emotions, EEG, Reading

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces—*Interaction styles - Gaze*; H.5.2 Information Interfaces and Presentation: User Interfaces—*Interaction styles - EEG*; H.5.4 Information Interfaces and Presentation: Hypertext/ Hypermedia

INTRODUCTION

Some texts are boring, some make us laugh, and some texts raise our interest. Unfortunately, almost all of these evoked emotions are lost. They are not recorded, let alone stored or searched for, except through user's manual interaction (rename *document* to *interesting document*). Even worse, labels given on a document level are crude and reduce the document's content to at most a few tags.

Also there is a number of web services like `slashdot.org` or `dailyme.com` that enable users to rate comments according to their evoked emotions. Currently, however, these services require manual interaction as well and, again, reduce the evoked emotions to a single statement.

Submitted for review to IUI 2011.

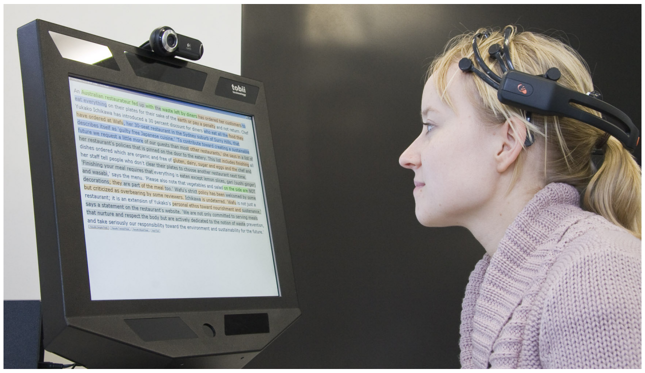


Figure 1. The experimental setup. The user sits in front of an eye tracking device while reading text. A web cam or neuroheadset track displayed emotions which are then annotated in the text.

In contrast to the websites and methods mentioned, we assume that the emotions experienced and displayed during reading are diverse. We acknowledge that one part of a text can be funny while another part of the same text may be sad, and thus we try to investigate what it would need to record and store these emotions in their whole complexity without any manual interaction.

Very little research has been conducted in the field of assigning emotions to text in real time. To date, a number of systems have been considering characteristic eye behaviors to recognize emotions while reading in real-time using an eye tracking device.They exploit variations in pupil size, blink rate and saccade length to identify the user's emotional state such as increased workload. Tiredness and attention are other emotions that can be detected by referring to changes in eye parameters [5]. However, emotions relying heavily on facial expressions such as joy and doubt cannot be detected by an eye tracker.

The section *Detecting Emotions* describes our methods used (also see Figure 1), in the section *From Emotions to Text* we report how we merged the detected emotions with eye tracking and in the *Evaluation*-part we report and discuss our findings. The paper concludes with an outlook on future work.

DETECTING EMOTIONS

The definition of emotions is rather complex. While some definitions are based on affective behaviors, others analyze cognitive and physiological reactions and again others combine all approaches and view emotions as an interaction between the mentioned factors[3]. We focus on a definition which considers what is externally *perceivable* and classify emotions according to the typically visible facial expressions. Thus, our working hypothesis is the existence of four commonly observable, mostly mutually exclusive emotions during reading: Joy, boredom, doubt and interest. These are chosen out of the belief that they are likewise of significance in human-document-interaction as well as that there is a reasonable chance for computer-aided recognition. With that in mind we investigate two methods to detect them.

First, we examine the use of a web cam mounted on the computer screen to record the reader's facial expressions while reading, see Figure 2. Every time a new image is captured, facial features like eyes and mouth are detected using Haar Cascade Classifiers in OpenCV. Eye and mouth regions are considered in particular because they characterize facial expressions uniquely. The located regions are extracted from the image and compared against an already classified training set of facial images by applying a trained SVM. The emotion of the closest match is then returned.

In our second approach, we replace the webcam by an Emotiv[1] neuroheadset. It measures facial expressions and fine muscular pulses to provide information about the subject's current emotional state. This muscular information can be combined with its tracking capability of brainwave signals, allowing for a detection of states even if they are not visible. We build upon the device's Emotiv API, which delivers processed EEG data in the form of different channels, each representing a subjective state or facial expression. We store all the brain specific data, and classify the emotions when needed, after taking the average values of each collected signal. The signals we consider for emotion classification are engagement (reflecting boredom and interest), furrowing (doubt), smiling and laughing (joy). We compare these values to the emotion characteristic values we collected in a calibration run and return the most likely emotion match. In case the measurements satisfy no known emotional pattern, a *neutral* state is returned, signifying that neither of the emotions is present.

Both methods have their strengths and weaknesses. While the webcam depends on visible factors, the neuroheadset has more of an *insight* into the human brain, providing more sophisticated information. During initial experiments we observed that, in contrast to the emotional expressiveness of human-to-human interaction (e.g., people laughing loudly about a joke told), the expressiveness when a reader interacts with text yields rather low levels of emotion (e.g., people may just show a faint smile if reading a joke). In that respect the neuroheadset is more sensitive since it detects also slight facial muscular movements.

The webcam on the other hand requires the emotion to be expressed in a high level of intensity in order to be detected in a satisfactory manner as shown in [2] with a low-cost webcam and in [4] with more sophisticated algorithms and apparatus. Also, the webcam does not allow any obstacle (e.g. a hand partially obscuring the reader's face) for the face to be analyzed, while the neuroheadset depends only on the direct connectivity between sensor and scalp to provide accurate brainwave data. However, the neuroheadset requires a more lengthy initial preparation, the sensors (and therefore the reader's scalp) have to be dampened and due to its applied pressure for sensor connectivity some users find it uncomfortable to wear after some time.

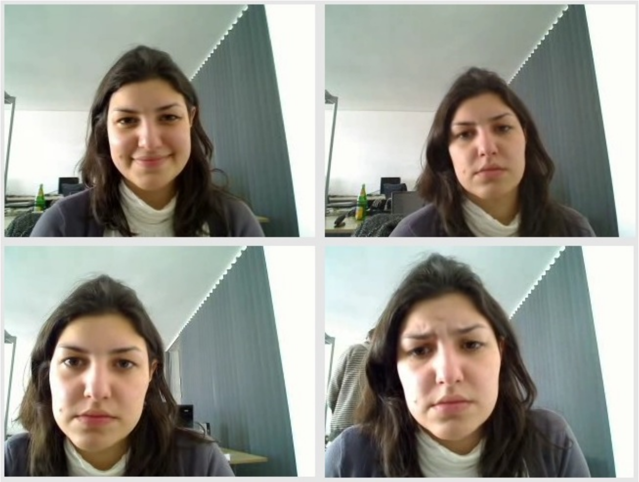


Figure 2. Samples captures of four emotions displayed during explicit web cam training. The four emotions are, clockwise starting from the upper left, joy, boredom, doubt and interest. It can be seen that boredom and interest are visually almost indistinguishable. In this respect EEG devices are likely to be of advantage as they are capable of detecting more subtle muscular activity as well as capable of evaluating brain waves.

Still, for the final prototype application, we preferred the EEG solution, as it provided more detailed emotional information and insight, specifically when emotions are not displayed visibly such as when reading.

FROM EMOTIONS TO TEXT

The rest of the setup consists of an eye tracker, a web browser and one of the emotion detection methods described above, compare Figure 1. The user opens an HTML document she wants to read in the browser, it is then loaded and segmented using the Text 2.0 framework[1]. After the page is fully initialized, evaluation of the gaze data and measured emotions begins.

The framework delivers fixation events to the document level and we use these to follow the reader's gaze over the text. By splitting the text into a set of span-elements, each containing a single word (compare [1] for details), raw fixation and emotion data can be assigned on word-level. This granularity allows for a detailed insight into the reader's emotional reading experience and it serves as a foundation to deduce the prevalent emotion on the paragraph or article level.

Each span element is augmented with a JavaScript `onGazeOut` event handler such that, whenever the reader's gaze leaves a word, a callback on that word is triggered. Next, the callback queries the present emotion and the currently focussed word is then tagged with an additional attribute `emotion`. The value assigned to it represents the emotional state that was evoked at this particular document position. At the same

[1] `http://www.emotiv.com/`

time the emotion values, along with the gaze data and the pages's structure, are written into a session log for later processing and retrieval.

In order to make sure that words are only augmented with emotions if they are actively read within their context and not only skimmed or unintentionally looked at, the global reading behavior needs to be kept track of as well. As long as the distance between the words read is within a certain bound the emotional tagging is active, otherwise the tagging stops until a *consistent* reading behavior is detected again. For the purpose of this study this was defined as reading any three of seven consecutive words in an incremental order, thus, from *left to right*.

If the reader moves his eyes back over previously considered words, i.e., when performing a regression, possibly different emotions are evoked than those detected during the first pass. In this case we do not overwrite the emotional information in order to keep the initial reading experience.

Also, while reading, not every single word in a sentence triggers an `onGazeOut` event, either due to normal reading behavior or eye tracking inaccuracies, leading to untagged words within a read sentence. So in case words are perceived but no fixations fall upon them, all the emotions evoked by those words will be assigned to the next element gazed upon. To avoid gaps between two fixations upon later retrieval and visualization emotions are also spread backwards. This means that on each new emotion assignment we proceed through the read text from back to front, and everytime an untagged word is encountered, it is assigned the emotion of the last word considered.

Interaction

In addition to the process of recording emotions we also investigated how they could be facilitated in real time. In this respect we followed the paradigm we already implemented in the Text 2.0 framework: gaze active handlers. In an initial approach we defined a set of attributes: `onSmile`, `onFurrow`, `onBoredom` and `onInterest`. These attributes can be used in parts of the web site's DOM tree. If then the eye tracker detects that the user's gaze is within the element's screen position and the emotion detection finds the corresponding emotions above a certain threshold the code within the handler is being executed. In this way, web designers can define event handlers for elements that are triggered if the user shows a specific emotion while looking at an element.

EVALUATION

We evaluated the overall system performance with respect to its overall accuracy of the tagged text. Based on a number of initial test runs comparing both emotion detectors we decided to use the EEG method throughout the experiment, as it proved to be less sensitive to head movement or rotation and provided more detailed insights in terms of sensor data.

The participants were seated in front of a desktop mounted Tobii X120 eye tracker and wore an Emotiv neuroheadset for the EEG emotion detection method as shown . Their task was to read five different articles in a browser. The articles were, for example, about bizarre scientific news, lengthy political discussions and funny jokes. Others were articles and threads retreived from `slashdot.com` and `dailyme.com` which were already classified by their users and we used this classification as a pre-selection to evenly distribute documents falling into different emotional classes. Upon completing each task, our users were then asked to mark the interesting, boring, doubtful, and joyful parts on the screen. Afterwards, the algorithmically computed emotions were compared against this feedback. Figure 3 shows a piece of text after tagging and coloring words according to the emotions evoked.

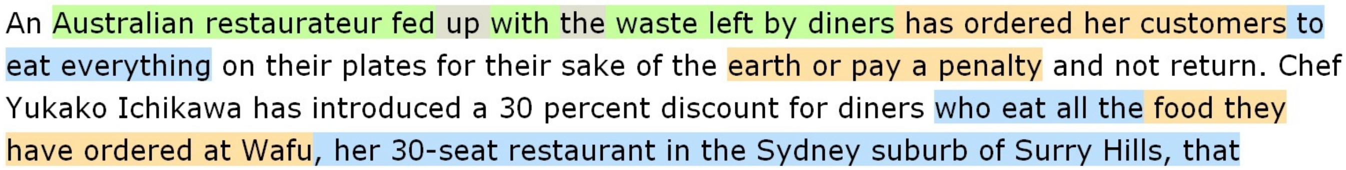
Emotion	Precision	Recall	F-Measure
Joy	0.74	0.93	0.82
Doubt	0.54	0.93	0.68
Interest	0.85	0.72	0.78
Boredom	0.67	0.13	0.22
Neutral	0.56	0.41	0.47

Table 1. The system performance with respect to the four emotions of interest. Joy and doubt are significantly better to detect than boredom.

In total, nine undergraduate students participated in the experiment and gave feedback on five articles. We considered 32 tagged texts for evaluation, 13 had to be discarded due to missing or low quality eye tracking or emotional tagging data. The emotions joy and doubt were evaluated on a sentence level, i.e. if a word was tagged with joy, although this particular word did not evoke this feeling, but another one in the same or adjoining sentence did, then it was considered as correctly classified and tagged. The neighboring sentences were allowed due to the fact that the emotions might be shifted because of the skipping or skimming of words.

The emotions interest, boredom and neutral on the other hand were evaluated on a paragraph level. This distinction was made because of the difference in the nature of the neuroheadset's signals. Joy and doubt depend on muscular movements represented by pulses and are usually instantaneously detected. The detection of interest and boredom is based on a continuous EEG data signal and it needs time to rise and fall with the reader's *mood*. Thus, since changes are not instantaneously detected, we agreed on a range of a paragraph which would provide enough time for the signal to stabilize itself and give correct feedback about the current emotional state of the user.

The results of this evaluation can be seen in Table 1. While the rather expressive emotions joy and doubt were often detected when they occurred, boredom was almost imperceivable. The most common cause for misclassification of joy and doubt were unintentional facial movements by the readers. This included moving lips while reading or furrowing the forehead when being highly concentrated. Based on the participants's oral feedback, we also found that the definitions of interest and boredom when related to text are ambiguous: Three participants mentioned that to them the opposite of an interesting reading experience was the neutral emotional state instead of boredom. And two participants defined an interesting text to be any text that handled a favoured topic, without considering their actual reading be-



An Australian restaurateur fed up with the waste left by diners has ordered her customers to eat everything on their plates for their sake of the earth or pay a penalty and not return. Chef Yukako Ichikawa has introduced a 30 percent discount for diners who eat all the food they have ordered at Wafu, her 30-seat restaurant in the Sydney suburb of Surry Hills, that

Figure 3. Sample image displaying a tagged text fragment and a number of issues. The participant read the text inside a browser while the emotions were recorded. After the reader finished, she was asked about her judgment with respect to her *real* emotions and the text was algorithmically colored. Blue denotes *interest*, orange stands for *joy*, green *doubt* and gray equals *neutral*. For the white parts no reading behavior was detected because the reader was laughing intensely which resulted in closed eyes and heavy body movement.

havior while going through the article. Thus, the evaluation results for these two emotions have to be considered as being rather coarse. It should also be noted that, due to the granularity applied, the reported number for precision and recall are likely higher than the device is actually capable of achieving in a frame-per-frame analysis of recorded emotions.

DISCUSSION

Assessing implicit emotions during reading is delicate. They are not as expressive as during human to human interaction, rendering them hard to detect. Likewise is their calibration and eventual evaluation, and many open questions remain.

Is text *interesting*, as one of our participants reported, when it adds new information about a topic which is of one's concern (while the text itself may be merely practical)? Or is it to be considered interesting only when one reads with a certain level of engagement and feeling of suspense? Does the *neutral* emotional state while reading exist, or is an article either interesting or boring? Should such emotions be defined subjectively and, if not, what would be a feasible set of emotions in terms of detectability and (semantical) expressiveness? How do we deal with emotions that take time to rise or decline and what are the sensible temporal and spacial limits of assigning them to text?

We also learned some lessons during our experiments. First of all, training and recording of emotions should happen implicitly. Recording explicitly evoked emotions might result in well divisible data sets, however, actual observable reactions show much less intensity than those displayed explicitly and thus, can not be categorized based on these. Also the generation of ground truth by having participants manually annotate their emotions, even after reading only a single document, proved to be unreliable as people often could not remember all evoked emotions anymore or were not sure about the specific place of evocation.

CONCLUSION & OUTLOOK

We presented a framework and a case study for a system that is capable of recording the emotional state of a user while reading text, and automatically assigns these detected emotions to the text. The evoked emotions and reading information can be used in real time or stored in a database for later retrieval.

Our results lead us to the conclusion that, given the emotion detection methods we employed, an algorithmic evaluation of emotions should happen well above word level. On this level a strict distinction into mutually exclusive emotions will also not hold anymore and should rather be expressed in terms of relative tendencies.

Another area in need of improvement are heuristics to deal with missing or inaccurate eye tracking data. For example, when laughing readers sometimes close their eyes or shake their bodies. In these cases neither can the eye tracker detect the readers eye and deliver fixation information, nor is a consistent reading behavior observed, thus resulting in missing emotional information about the text.

For the future we anticipate a number of applications where the emotional tagging system could be integrated to enhance the user experience by emotionally interacting with text. Possible applications include searching for text parts or articles that evoked certain emotions either by the reader himself while reading an own document or emotions which were displayed by other readers as a form of emotional feedback. Authors could also make use of the automatic real-time tagging and retrieve information about how readers respond to writings and help them analyze texts emotionally. Besides the emotional tagging, applications could actively interact with the user and react to the emotions displayed by using the emotional event listeners introduced to the Text 2.0 framework[1]. Finally, the tagging system can be easily expanded such that it can be applied on non-textual elements such as images and videos as well.

REFERENCES

1. R. Biedert, G. Buscher, S. Schwarz, M. Moeller, A. Dengel, and T. Lottermann. The Text 2.0 Framework. Presented at International IUI 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction, 2010.

2. E. Cerezo, I. Hupont, C. Manresa-Yee, X. Varona, S. Baldassarri, F. J. P. Lpez, and F. J. Sern. Real-Time Facial Expression Recognition for Natural Interaction. In J. Mart, J.-M. Bened, A. M. Mendona, and J. Serrat, editors, *IbPRIA (2)*, volume 4478 of *Lecture Notes in Computer Science*, pages 40–47. Springer, 2007.

3. P. R. Kleinginna and A. M. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, 5(4):345–379, December 1981.

4. M. Pantic and L. J. M. Rothkrantz. Automatic Analysis of Facial Expressions: The State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1424–1445, 2000.

5. M. Porta. Implementing Eye-Based User-Aware E-Learning. In *CHI 2008 Proceedings*, pages 3087–3092, 2008.