Night Mode, Dark Thoughts: Background Color Influences the Perceived Sentiment of Chat Messages

Diana Löffler(✉), Lennart Giron, and Jörn Hurtienne

University of Würzburg, Würzburg, Germany
{diana.loeffler, joern.hurtienne}@uni-wuerzburg.de,
lenn.geh@gmail.com

Abstract. The discussion of color in HCI often remains restricted to issues of legibility, aesthetics or color preferences. Little attention has been given to the emotional and semantic effects of color on digital content. At the example of black and white, this paper reviews previous studies in psychology and reports an experiment that investigates the influence of black, white and gray user interface backgrounds on the perception of sentiment in chat messages on a social media platform (*Twitch.tv*). Of sixty-seven participants, those who rated the messages against a black background perceived them more negatively than those who worked against a white background. The results suggest that user sentiment perception can be influenced by interface color, especially for ambiguous textual content laced with irony and sarcasm. We claim that this knowledge can be applied in persuasive interaction and user experience design across the entirety of the digital landscape.

Keywords: Color · Affective bias · Sentiment analysis · User interface design · Online chat · Embodiment · Conceptual metaphor theory · Persuasive computing

1 Introduction

Dark themed user interfaces, night mode, low-power state - light and darkness are fundamental experiences in our very (digital) lives, both literally and metaphorically. As diurnal animals, humans are active at daytime, when their visual sense works best, and inactive during the night. Unsurprisingly, *black* and *white* are the first basic color terms that enter languages around the globe [5]. Interestingly, black and white also carry evaluative and affective metaphorical meaning. We talk about a *black cat* crossing one's path as a sign of bad luck. A *black sheep* represents a bad entity in an otherwise respectable group, while *blackmail* carries malicious intent. When we ask someone to *keep it dark*, we ask for his or her confidentiality. Across many cultures, the color *black* has unpleasant connotations, involving death, evil, the void, and secrecy. On the contrary, *white* is associated with innocence, benevolence, purity and the divine: A *white lie* is harmless or even beneficial in the long run. A *bright future* conveys hope, and a *bright*

R. Bernhaupt et al. (Eds.): INTERACT 2017, Part II, LNCS 10514, pp. 184–201, 2017.
DOI: 10.1007/978-3-319-67684-5_12

girl is clever. Expressions of *white* and *black* relating to good and evil persist across an overwhelming majority of communities and cultures. Also, it has been shown that such metaphorical expressions are not mere figures of speech but instead reflect the underlying mental models of the speaker. As stated in Conceptual Metaphor Theory, many linguistic metaphors indicate how we think about the concepts linked by the process of metaphor [33, 34], and this approach can be applied to color psychology [36].

The dichotomous nature of *black* and *white* as opposites lends itself to facilitate the mental model of more abstract concepts like valence or morality, and has received support in cognitive linguistics as well as by empirical studies in the social sciences. For example, basic research in psychology demonstrated that the perception of *bright* vs. *dark* font color has an impact on the semantic categorization of affectively valenced word content. With 980 participants, diverse in terms of race, age, and geographical location, Meier and colleagues showed that response times in a two-alternative forced choice task are facilitated when word valence metaphorically matched font color (i.e., a positive word written in *white* font color) compared to when they did not match (i.e., a positive word written in black font color) [38].

Naturally, we also rely on color information when making inferences about digital information. As the personal computer is foremost a visual medium, the same effects may be replicable in the context of Human-Computer Interaction (HCI). Digital content is primarily conveyed via text, images, and video, all of which are realized by displaying contrast and color gradients to transmit information. Thus, HCI systems are inherently prone to subjecting users to cognitive biases caused by the perception and subconscious processing of black and white. So far, however, design recommendations involving color are almost exclusively restricted to issues of legibility, aesthetics or color preferences [37]. Comparably little attention has been given as to how it affects us emotionally and how it might influence the semantic interpretation of digital content and even interpersonal exchange [36].

To this extent, digitally mediated human communication is an intriguing area. Contrary to face-to-face communication, the clear majority of exchange in social media is conducted via textual messaging, stripped of many nonverbal signals, thereby increasing the ambiguity of the content. Messages with room for interpretation are particularly susceptible to incidental environmental cues [44]. The influence of black and white effects could have a significant impact on the judgment and interpretation of unclear, ironic or sarcastic messages. Digital chat rooms involving two or more users, a format still widely used and more prevalent than ever[1], bear potential for cognitive biases, as they require the users to engage from remote terminals with varying form factors and environmental conditions. This imposes the question if heterogeneous chat interfaces provide different messaging environments to such an extent that contents may be interpreted divergently on a semantic and affective level.

[1] (2017, January). Most popular mobile messaging apps worldwide as of April 2016, based on number of monthly active users (in millions), retrieved from https://www.statista.com/statistics/258749/most-popular-global-mobile-messengerapps/.

2 Related Work

2.1 Basic Research in Psychology on Affective Biases Related to Black and White

Much work has been done in perception psychology to investigate fundamental effects and biases when processing visual information of varying monochrome contrasts. For instance, Chiou and Cheng [11] employed an altered version of the classical Stroop-task [46] to study subconscious moral connotations of perceived font color (black vs. white). In the Stroop-task, participants are asked to name the font color of a task-irrelevant but color-relevant word. The closer the semantic relationship between word and font color, the higher the observed interference effect [27]. Chiou and Cheng found that the reaction time for color naming was faster when words related to immorality (rather than morality) were written in *black* (e.g. "malevolence"), and when words related to morality (rather than immorality) were written in *white*. This shows that although the semantic content of the to-be-classified words was irrelevant for the task, subjects nevertheless processed it and word meaning interacted with the primary task of font color categorization.

Similar effects could also be shown in other experimental setups. For example, when subjects are exposed to different levels of environmental illumination, the salience of moral considerations and the likelihood of engaging in ethical behavior is increased when the level of illumination is increased [11]. These results suggest that brightness might increase the mental accessibility of ethical behaviors. This affective bias induced by the perception of brightness could also be replicated in the context of purely visual information. Lakens, Fockenberg, Lemmens, Ham and Midden fairly recently discovered that brightness heavily influences the cognitive and emotional evaluation of affective pictures [31]. Over the course of multiple experimental setups, brighter pictures were evaluated more positively, whereas darker pictures received more negative judgment. Therefore, the perception of brightness seems to bias affective and moral interpretation, regardless of the conveying medium.

Another field of investigation is the directionality of the relationship between brightness and valence. In a study by Meier and colleagues, subjects' perception of gray-gradients was biased after being presented with morally connotated words [39]. Moral and ethical words, like *innocence, purity* and the like caused participants to judge different gray-gradients to be brighter. Similarly, when immoral terms were presented, participants subjected darker ratings, even though the colors were identical in both conditions. Findings like these suggest that the link between brightness and valence is bidirectional.

Some scholars focused on determining if the apparent link between brightness and valence is automatic. Meier et al.'s work indicates that these affective mappings are automatic in nature and thus elude conscious control without proper training [39]. On the contrary, Lakens, Semin and Foroni questioned the automaticity of the mapping and conducted a series of six experiments on the subject matter. Their experimental paradigm aimed at removing the factor of linguistic idiosyncrasy of one's native language by presenting the test subjects with foreign Chinese characters [32]. The Chinese alphabet is made up of ideograms, stylized depictions of superordinate ideas and

concepts – like pictograms. For example, a vertical brush stroke with a horizontal line intersected with two diagonal lines branching from the top down and outwards constitute the word tree. Simplified, it is a symbolic depiction of a stem with branching twigs. Two of these "tree" characters side by side add up to the word forest. Lakens et al. instructed participants, devoid of any understanding of the Chinese language, to judge whether a correct translation for a character was presented on the screen while manipulating the font color (*black* vs. *white*). The authors found that when *white* font color was shown together with a positive translation and *black* font color with a negative translation, the correctness of the judgments was above chance. This indicates that the effect of perceiving *black* vs. *white* transcends peculiarities of individual languages and linguistic expressions, but instead represents the underlying conceptual metaphor GOOD IS BRIGHT – BAD IS DARK [3]. However, *white* only evoked positive associations when the negativity of *black* was co-activated but remained neutral when perceived alone. Thus, GOOD IS WHITE seems to be more context dependent that the association BAD IS BLACK.

Consequences of such cognitive biases are often reflected in behavior and decision making. Exposure to incidental light and dark visual contrasts tends to lead people to think in a "black and white" manner, as more extreme moral judgments follow than those which would occur otherwise [48]. Intuitive processes are affected in a way that leads to a polarization of moral judgment. Darkness appears to encourage moral transgressions. Zhong, Bohns and Gino found that it seems to induce a state of illusory anonymity which promotes egocentric behavior and dishonesty [49]. In their study, participants wearing sunglasses behaved more selfishly and cheated more often and systematically. This effect appears to scale from behavioral trivialities to severe crime. Regardless of factual anonymity, a psychological effect kicks in, suggesting impunity and a lack of repercussions, as darkness can conceal identity and thus legal pursuit. This is apparent in criminal statistics, as illumination directly correlates to the number and severity of criminal incidents [17]. In this vein, improved street lighting can have an immediate preventative effect on crime.

Similarly, the color *black* has been linked to behavioral aggression. In their classical study, Frank and Gilovich explored the effect of *black* uniforms and jerseys in professional sports in relation to the number of caused penalties. They found that throughout the course of their observations, teams wearing *black* jerseys constantly ranked among the top of the penalty statistics [19]. In the event a team switched from a *non-black* to a *black* uniform, an immediate increase in penalties was noticeable. Frank and Gilovich inferred that *black* uniforms lead to both biased judgments of referees and increased aggressiveness of the players themselves, which can be attributed to social perception and self-perception processes.

2.2 Affective Biases of Color Perception in HCI

In the field of HCI, the impact of colors on human behavior should be considered when designing User Interfaces (UIs). Predominantly, related work in this area is focused on the perception and effects of chromatic colors and their affective impact. For example, Hawlitschek, Jansen, Lux, Teubner and Weinhardt linked UI background color to reciprocation behavior via perceived warmth and color appeal [24]. In electronic

commerce, manipulation of the affective state of mind is a potent tool in guiding reciprocity behavior and commercial purchasing power. The invocation of hedonistic notions like *desire*, *need* and *urge* regulates buying intentions [40]. This emotional reaction can be achieved by evoking and enhancing the customer's pleasure. According to Allagui and Lemoine, proficient color design increases pleasure, whereas inappropriate use of colors results in boredom, which in turn negatively effects mood and lessens the User Experience, leading to a significant decline of affective buying impulse [2].

Within the confines of digital chatting, color has widely been instrumentalized as an indicator of human sentiment and mood. *Sentiment* is defined as the attitude or state of mind toward or induced by something (contrasting to *mood*, which is defined as the affective state of mind lasting longer than mere emotions) [14, 15]. Together with emoticons, color provides the primary means of expressing and conveying affective states. For instance, Dos Santos, Gestraud, and Texier filed a patent for dynamically evaluating the mood of a chat user as a function of color and its intensity [43]. Back in 2006, Sánchez et al. suggested an instant messaging system tailored to convey mood and emotion [42]. Besides employing emoticons, colorful bubbles encompassing each message and font size maintain a record of current and past moods. An idea that has since been implemented in the globally popular Facebook Messenger[2], which allows for color personalization of any individual chat dialog or -group.

Since chromatic effects are so potent and powerful in guiding user behavior, comparatively little attention has been given to the domain of affective studies based on the sole perception of black and white in digital environments. Research in this domain is virtually limited to the topics of eye-strain, ergonomics (e.g. [29]) and workflow efficiency and -optimization (e.g. [10]). Although the presence of affective biases attributed to black and white in the digital context appears highly likely, validation studies are noticeably small in numbers [16]. This research gap is especially lamentable because perceptive *black/white* effects seem to be remarkably robust and replicable [38].

2.3 Automated Sentiment Analysis

Sentiment and mood have long been a growing area of interest in computer science. Today, the extraction of human sentiment and opinions from Big Data is one of the foremost researched topics [7, 8], involving data mining and using Artificial Intelligence and Machine Learning (especially Deep Learning), bearing immense potential for applications, both scientifically and commercially. As it sheds light on user desires, political and domestic satisfaction, as well as any other marketable area of life on a mass scale, the computational treatment of sentiment and subjectivity has received extensive attention. In the past years, major advancements in automated sentiment extraction have been made that enable a potent approximation of actual affective states conveyed in chats, blogs, social networks and other opinionated outlets.

For instance, a new approach to *Natural Language Processing* enabled Kouloumpis, Wilson, and Moore to analyze the sentiment on the popular micro-blogging service *Twitter* by identifying part-of-speech features combined with the presence of

[2] (2017, January). https://www.messenger.com/.

intensifiers, like emoticons and common abbreviations [28]. Through multiple iterations of training data, they could achieve an accuracy of upwards to 75% for classifying *Tweets* into the categories "positive", "neutral" and "negative", measured against a human evaluation. Similarly, Godbole, Srinivasaiah, and Skiena developed a system for large-scale sentiment analysis over large corpora of news and blogs [22]. By assigning complexly computed scores, indicating positive or negative sentiment, this system can describe the mood within any given text with each entity's sentiment relative to other entities of the same class. This allows for spacial analyses and distribution of opinions and feelings to provide sentiment maps and thus give a comprehensive overview of a certain issue over large amounts of users. Given a certain robustness and accuracy, sentiment extraction algorithms could function as an indicator of ‚objective' chat sentiment, independent of particular UIs and their inherent tendency to inflict cognitive and affective biases.

2.4 Research Question and Scope of the Empirical Study

In basic psychological research, the influence of *light* and *dark* on affective states has been thoroughly examined. It has been validated that the perception of *black*, *white* and chromatic colors ubiquitously and unconsciously takes part in governing cognition, emotion and behavior. Furthermore, automated sentiment extraction receives growing interest and scientific exploration in recent years. However, extensive sighting of the related literature landscape revealed a blind spot of validation studies dedicated to the reproduction of basic research on affective biases mediated by the perception of light and dark in the context of HCI. Since the manipulation of color and contrast in digital environments is one of the most trivial alterations to be made, research on the ensuing effect on perceived sentiment and mood is highly warranted.

Legibility on computer screens is superior when using *dark* characters on a *light* background. Alongside numerous studies coming to this conclusion, Bauer and Cavonius found that participants were 26% more accurate and significantly faster compared to reading from an inverse color scheme [4]. Conversely, with *light* text on *dark* backgrounds, the iris opens wide to let in more light and causes a slight deformation of the lens, resulting in blurred letters which are marginally harder to perceive. This effect is enhanced for people with astigmatism. However, especially near night-time, prolonged exposure to *bright* computer screens negatively affects sleep, circadian cycles and is known to cause headaches [9]. Consequently, *dark* themes minimizing the presence of *bright* elements are widely used amongst professions that heavily rely on working on computer screens. Additionally, digital services all over the internet employ *dark themes* and interfaces for purely cosmetic reasons. The sheer number of possibly affected people worldwide motivates and merits the investigation of psychological effects of perceiving *light* or *dark*, *black* or *white* in the context of HCI.

Human communication lanced with irony and sarcasm provides ample grounds for misinterpretation of affective intent and entails an enormous potential scope of severe consequences, both on an interpersonal and professional level. Today, decades after its inception, chatting persists as the primary means of digital communication globally. Thus, this work aims at investigating the influence of a predominantly *white* interface on perceived chat sentiment in contrast to a primarily *black* interface. It is expected that

when participants in an experimental study have to evaluate the sentiment of a chat excerpt with ambiguous messages, their judgments will be more negative when the background color is *black* compared to when the evaluation is performed in a predominantly *white* interface. If the interface color is gray or the sentiment analysis performed automatically, the judgments should be less extreme, as they are not influenced by a valence-related background color of the UI. Moreover, we argue that the expected effect is driven by the UI background color as opposed to the font color of presented messages, because the participants had to take the whole interface into account (video, chat log, chat message, answer keys) and the background is more prominent in the visual field of the subjects compared to most psychological studies that only present single words against a gray background, e.g. [38]. In such forced-choice reaction time tasks, the participants' full attention is directed to the target stimuli, thus inverting the prediction. Such inverted color effects are typically found in color psychology when single-color evaluations are extended to more applied contexts and interactions [25], thus questioning the transferability of results of basic research to more applied contexts.

3 Method

3.1 Participants

Test Subjects. Published effect sizes of related basic research in psychology were injected into *G*Power* [18], a statistical power analysis tool, to compute an approximation of required test subjects to find comparable effects in this study. These calculations indicated a sample size of 60 to 70 participants. The subjects were recruited through an online recruiting system in exchange for course credit. They were screened to meet *Twitch.tv's* target demographics of (predominantly male) 18 to 35 years old gaming savvy users. The total sample of 67 participants consisted of 24 women and 43 men aged 19 to 29 years (M = 21.167, SD = 1.932).

Expert Group. Six 'expert' evaluators could be won to participate in the study as a reference to the experimental groups. Subjects from this group were deeply involved with the subject matter of *Twitch.tv* streaming as a part of their professional occupation at a company situated in the social streaming market, as well as their everyday private lives. Due to their profound knowledge of vernacular and game mechanics, they are highly capable of disambiguating messages semantically with an outstanding sense of context and intention of the original author. This leaves less room for interpretational errors and approximates the true chat sentiment more closely, than the large group of test subjects. The expert sample consisted of 6 men aged 25 to 29 years (M = 25.83, SD = 1.46).

3.2 Procedure

In a single-blind between-subjects design, test subjects were assigned to either be part of condition white, processing the chat messages against a white background color, or

condition black, where they encountered the black version of the UI (counterbalanced across both conditions). The study was set up as an online experiment that participants accessed through an email link. They received written instruction to rate the sentiment of presented chat messages based on their subjective perception and as fast as possible to emulate authentic reactions while reading a chat. To prevent participants from guessing the purpose of the experiment and thus falsifying resulting data, they were given a misdirecting study title, which was clarified upon completion. After completing the sentiment evaluation of 229 chat messages by labeling them as "positive" (+1), "neutral" (0) or "negative" (−1), an average sentiment score was calculated for each participant. In case a participant disregarded or missed a message, the denominator (i.e. total number of messages) was adjusted accordingly to produce a sound average. After having worked through all messages, a demographic questionnaire had to be completed with age, gender, average internet usage, gaming affinity, *Twitch.tv* familiarity, schadenfreude, mood, and tiredness. The whole procedure lasted 45–60 min.

As a reference measure for the experimental groups, two statistics have been employed. Firstly, another variant of the interface was generated in neutral gray [1, 21]. On this version, independently from the experimental conditions white and black, six *Twitch.tv* experts labeled the same chat excerpt as the experimental groups. Secondly, to access chat sentiment more objectively, a sentiment analysis algorithm was applied, eliminating human factors and thus susceptibility to cognitive biases and fatigue (see paragraph 3.1 Sentiment analysis algorithm).

3.3 Material

Chat Excerpt. Founded in June 2011, *Twitch.tv* has set out to become the world's leading social video platform and community for gamers. As of the time of this work, it is the largest contemporary facilitator of chat messages with an average volume of 200 billion messages per day[3]. *Twitch.tv* employs a chat with often hundreds or thousands of simultaneous participants. The publicly available nature of these open chat rooms allows for an easier approach to recording and logging prolonged exchanges of chat messages. This feature, combined with the fact that *Twitch.tv's* sheer throughput dwarfs that of all of its competitors, has led to the decision to base this study on the example of *Twitch.tv*. By employing *Chatworkers* (bots that sit inside a specific channel and record the chat log), any arbitrary sequence of messages in any channel on *Twitch.tv* became obtainable. Moreover, *Twitch.tv's* broadcasts and chats are accessible to the public and recorded and archived by default to be readily available as videos on demand for certain periods of time.

As eSports (professionally organized and managed competitive gaming) is driving the highest traffic per channel on *Twitch.tv*, we chose a chat excerpt with representative qualities for this format. The choice fell on a broadcast of a competitive contest of one of the most popular eSports games, DOTA 2, namely a match featuring competing teams *Natus Vincere* vs. *No Diggity* from *Dreamleague Season 5*, which aired 18:30 CEST on

[3] (2017, January). Twitch.tv message throughput. Retrieved from https://jobs.lever.co/twitch/61fb0c31-6f59-435b-a4f7-ea0e57a038b6.

April 05, 2016 on *Twitch.tv*. Recorded in its entirety, the encounter resulted in approximately two hours of video with its corresponding chat transcript.

When cropping down the recorded material to attain a prolific base of chat data for participants to evaluate, a segment of the recording was chosen that contained both a negative and a positive event, as well as the conclusion of the match, leading to mixed verdicts and opinions about performances throughout the course of the encounter. The chat excerpt accompanying the chosen content sequence, spanning the time of 10 min, originally consisted of 1172 messages. Early testing indicated a completion time of roughly 22 min per 200 labeled messages. To achieve an overall completion time of about an hour to avoid user fatigue, the chat log was trimmed down to a total of 229 unique messages and aliases were removed. Using a basic JSON script, the original log was processed to remove overly crude profanity, redundant messages, as well as messages with less than three characters. Although the excerpt was shortened drastically, much care was taken to preserve the chat's representative qualities as to prevent contortion of the content and flow due to overly manipulating the source material.

Evaluation Interface. To provide a functional environment for sentiment evaluation, video and condensed chat log were fed into a browser-based interface based on HTML/CSS framework Bootstrap. Messages could either be rated "negative" (Hotkey A), "neutral" (Hotkey S), "positive" (Hotkey D) or "undefined" (Hotkey X), with the latter being reserved for messages that by nature cannot be interpreted properly by sentiment. Examples for this case include Cyrillic characters, ad- and spam links, as well as a language other than English. Above the message field, the accompanying video scene was embedded. Initiated by triggering the "next" button, the clip would play back in real time until reaching the consecutive message of the processed chat log. From this labeling interface, three different versions were created. The *dark* version featured a *black* background (hex code #000000) with *white* font and button borders. Conversely, the *light* version consisted of a *white* background (hex code #FFFFFF) with *black* borders and font (see Fig. 1). For the expert group, a *gray* variant (hex code #7F7F7F) of the interface was generated. On the back-end, information for the labeling of any message was stored on a deployment of open source, multi-model database distribution *Couchbase*.

Auxiliary Data. The accompanying demographic questionnaire inquired information about gender, age and a range of possible moderator variables. Average daily internet usage, gaming affinity, and weekly *Twitch.tv* usage are likely to influence sentiment labelings of test subjects since they largely constitute a certain knowledge base of internet culture and language. Internet- and *Twitch.tv* usage data were gathered on a five-point Likert scale, ranging up to more than five hours per week.

Gaming affinity was queried on a five-point Likert scale ranging from "not familiar at all" to "very familiar". Furthermore, another factor which can potentially influence the evaluation of sentiment is the trait of *schadenfreude*, the pleasure at the suffering of others [45]. Depending on the character, a participant might find an insult, witnessed in the chat or on the video feed, either funny and gloat over someone's misfortune or offensive leading to rejection. To determine the experimental subjects' spitefulness and taste in humor, they were presented with three different insults that occurred in the captured broadcast before the segment chosen for the actual experiment. They were

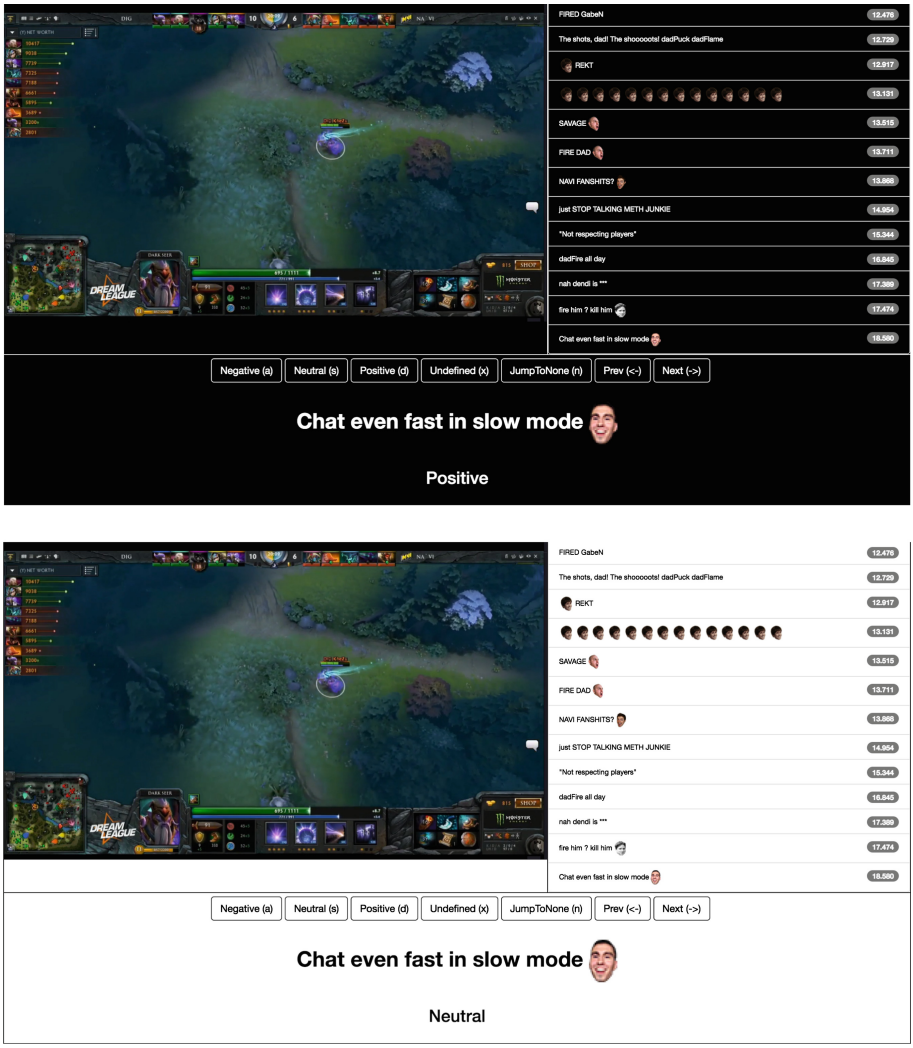

Fig. 1. Top: black interface variant. Bottom: white interface variant, with an exemplary message and Twitch.tv face emoji.

asked to assess whether they find these messages to be funny or rather offensive. Finally, the participants' condition and state of mind before conducting the labeling were inquired. Mood and tiredness, evaluated with four ordinal polytomous options, derived from the Epworth Sleepiness Scale [26], were inquired.

Sentiment Analysis Algorithm. Originally conceived to analyze sentiment in live Tweets, SentiStorm [47] was adapted to learn and interpret language specific to the platform *Twitch.tv*. The algorithm broke down every message word for word into tokens. Each token possesses a feature vector, containing 500 attributes. These attributes

can be simple classifications (noun, verb, adjective) or more complex linguistic constructs, like *term frequency* or *inverse document frequency* – both related to language- and document occurrence of a word. Individual features were then equally weighted with a value between "0" and "1" and, in turn, produce an average numerical value from "−1" to "1". These values are then fed into a *Support Vector Machine*, that calculated in which category a message falls into, measured relatively against all other messages. For this to work, the algorithm had to be trained. For this purpose, 2000 messages from the stream broadcast before the experimental excerpt was used.

4 Results

4.1 Chat Sentiment

The impact of background color as the independent variable on sentiment ratings of the participants was tested using a one-way ANOVA. Alpha was set at .05. The results indicate a significant influence of background color on sentiment ratings, $F(2, 72) = 7.994$, $p = .001$, $\eta^2 = .186$. A Tukey post hoc test revealed that the sentiment ratings were significantly more negative when participants worked on a *black* background (−0.255 ± 0.441) compared to a *white* background (−0.069 ± 0.069), $p = .001$, *Cohen's d* = .958. There were no statistically significant differences between the white and black background color conditions and the experts working against a gray background, $p = .928$ and $p = .178$, respectively. The *Twitch.tv*-experts judged the chat sentiment to be situated between the numerical values of the experimental conditions (−0.101 ± 0.147), see Fig. 2. Next, the three participant groups were compared against the computed mean value of the employed sentiment algorithm of $M = -0.013$, using three one-sample *t*-tests. While the participants in the *black* background condition rated sentiment significantly more negative than the algorithm, $t(33) = 7.233$, $p < .001$, $d = 2.557$, the values of the *white* and *gray* background condition did not differ statistically significant from the algorithm, $p = .111$ and $p = .185$, respectively. Overall, the average values in all conditions were negative.

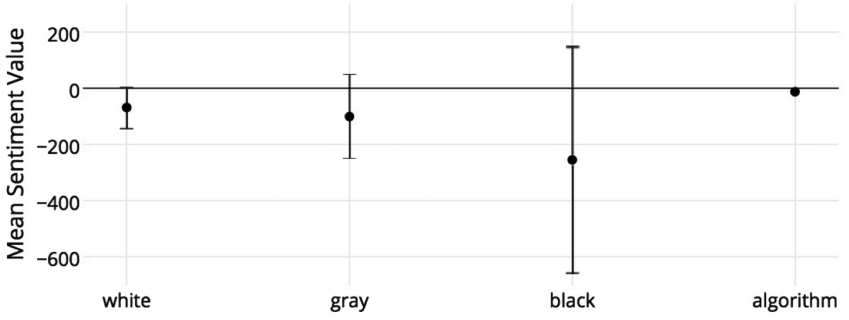

Fig. 2. Averaged sentiment judgments (−1 to 1) of the two experimental conditions (*black* vs. *white* background), expert group on *gray* background and algorithm; error bars, if applicable, indicate 95% confidence intervals.

4.2 Control Variables

As both experimental groups did not differ from each other regarding the collected control variables, values are reported for the whole group of test subjects. On average, participants stated that they spend three to four hours utilizing the internet daily. On a five-level Likert scale, ranging from "not familiar at all" to "very familiar", participants rated their familiarity with gaming contents as "familiar" ($M = 3.933$, $SD = 1.913$), and with eSports contents as "moderately familiar" ($M = 3.100$, $SD = 1.191$). Furthermore, participants rated their current mood before the experiment as "neutral" to "positive" ($M = 3.560$, $SD = 0.633$) on a second five-point scale. On average, participants rated their sleepiness on a four-point one-item scale, derived from the Epworth Sleepiness Scale and ranging from "would never doze" to "high chance of dozing", between categories "would never doze" and "slight chance of dozing" ($M = 1.950$, $SD = 0.746$). Qualitatively, weekly *Twitch.tv* usage averaged out to two to three hours on an ordinal categorical scale. Concerning schadenfreude, 26 participants declared the presented insults as being funny, whereas 34 found them to be offensive ($M = 1.567$, $SD = 0.500$, with the value "1" being coded for "funny" and "2" meaning "offensive").

To estimate the relative impact of the assessed control variables on the sentiment judgment of the participants, a multiple linear regression analysis was performed with interface color, sex, age, internet use, gaming-, eSports and *Twitch.tv* affinity, mood, tiredness and schadenfreude as independent variables and sentiment rating as dependent variable. The ten variables explain a significant amount of variance in sentiment ratings, $F(10, 59) = 3.295$, $p = .002$, $R^2 = .402$, *adj.* $R^2 = .280$. However, only interface color ($\beta = -.393$, $p = .002$) and schadenfreude ($\beta = -.436$, $p = .001$) were significant predictors of sentiment rating. If the interface background color was *white* and if a participant perceived written insults rather funny than offensive, the sentiment rating was more positive.

5 Discussion

This work set out to investigate the influence of *light* and *dark* chat interfaces on perceived sentiment. The resulting data suggests the presence of such affective biases, explored and validated in basic psychological research, in a digital environment. Sentiment ratings of participants in the *black* background condition were significantly more negative than those of participants in the *white* background condition, indicating affectively biased perception of mood in the chat excerpt. One alternative explanation for this result is that the participants' decisions were biased by differences in perceptual fluency. As *white* text on a *black* background is more difficult to read than *black* text on *white*, the perceptual fluency is lower, causing more negative judgments of the content [41]. Comparing the foreground/background contrast of both interfaces after the Web Content Accessibility Guidelines 2.0 using a tool provided by Hülsermann[4], no differences in contrast ratio and readability could be found. Moreover, the readability of the *gray* variant was notably lower than the *white* and *black* version, but experts did not

[4] (2017, January). http://joerghuelsermann.de/tool/kontrastrechner.

subject the lowest affective ratings. Therefore, this alternative explanation seems less likely. A second alternative explanation is the impact of familiarity, as more familiar things are judged more positive [20]. As the most consumed text is black font on white background, this could lead to a familiarity effect which could have impacted the interpretation of texts to a more positive sentiment of the white background. However, we could find similar ratings for white and gray (mid-gray is certainly less familiar than white background color). In addition, the black sentiment judgements were constantly negative over time (45–60 min exposure during the experiment), and a mere exposure effect (repeatedly perceiving things results in a positive bias) could not be observed. Therefore, we conclude that this does not offer a better explanation of the results than the brightness-induced affective bias hypothesis.

Although the sample size of the *Twitch.tv* expert group was small, their judgments, unbiased from the background color, do provide a frame of reference and valuable insight into more precise judgments of the inherent sentiment of the sample excerpt. Their cumulative ratings were situated numerically between those of participants using the *white* interface and those using the *black* interface. Statistical tests showed that while the *white* interface did not lead participants to rate the chat messages more positively, perceiving them against a *black* background dramatically decreased the perceived chat sentiment. This is in line with the literature that *black* automatically evokes negative associations, whereas *white* remains neutral when the negativity of *black* is not co-activated [32]. It is a subject to further research whether perceiving chat messages against a *white* background increases the perceived positive sentiment when a within-subjects design would have been employed. Moreover, future studies could employ a control condition that avoids an influence of color overall, for example by using an audio condition as baseline, and control for surrounding light sources.

Overall, the sentiment in the chat has been classified as slightly negative, as all average values are below "0". This fact has been anticipated since *Twitch.tv* is known to be somewhat notorious for a crude form of manners and verbal inconsideration amongst users.

To quantify inter-rater agreement across all messages, Fleiss' Kappa κ has been computed. In its unadjusted form, containing missing ratings where participants had omitted single messages and resulting in uneven counts of raters, the agreement was fair ($\kappa = .509$, $SE = .0014$), according to [35]. Standardized to 67 raters for each message, the measure of accordance fell to $\kappa = .198$, $SE = .001$. Over such many messages, and accounting for the highly subjective nature of mood perception, a low consensus is to be expected.

The distribution of age ($M = 21.167$, $SD = 1.932$) and sex (24 women to 43 men) of participants were within the intended demographic cohort, adding to the validity of gathered data about the factual user base of *Twitch.tv*[5] of 75% Millenial males. Of the collected control variables (age, sex, internet use, gaming-, eSports and *Twitch.tv* affinity, mood, tiredness and schadenfreude), only schadenfreude had a significant impact on sentiment ratings. The more the participants interpret insults against other people as humorous rather than offensive, the higher the perceived chat sentiment.

[5] (2017, January). http://twitchadvertising.tv/audience/.

Decreased empathy in online environments [6] together with exposure to very competitive surroundings, as is the case in eSports, likely explains spiteful and gloating tendencies like schadenfreude and in turn its effect on perceived chat sentiment [23].

Participants' mood before the experiment was rated "neutral" tending towards "positive", which is consequential when assessing affective sentiment, as one's state of mind will always factor into perception. Fortunately, the effects of potentially confounding factors mood and sleepiness on sentiment ratings were negligible in our data, as no significant contribution in the regression model could be found.

The experimental manipulation resulted in a decisively large effect size according to Cohen, signaling a strong statistical difference between the experimental conditions of different UI background color [13]. The employed ordinal scale for measuring sentiment with merely three different levels ("positive", "neutral", "negative") might have contributed to the strength of the effect. However, since only the control variable of schadenfreude had a slight impact on sentiment ratings and considering the random assignment of participants to the interface conditions, a potent and compelling effect on perceived sentiment attributed to the interface can be inferred.

Several limitations of this study need to be considered. First, the suitability of experimental subjects was not thoroughly validated. Although during recruitment the call for participation postulated gaming and eSports knowledge and prolific understanding of the English (internet-)language as prerequisites for attending this study, it is questionable if all participants truly possessed a thorough knowledge of gaming-related terms, idioms, and mechanics despite claiming to do so. A broader study in this endeavor should have employed a test to assess the gaming-savviness of participants.

Second, concerning the experimental design, a more finely grained scale for measuring sentiment may have yielded more revealing results. However, affective interpretation occurs automatic and instantaneously. If participants ponder too long to accurately assess the mood on an elaborate scale, the authentic replication of the actual human process of sentimental perception could be problematic. Additionally, in respect of a large number of to-be-evaluated messages, this might result in participant exhaustion and in turn produce unnecessary noise in the resulting data. A balance between authenticity and information entropy would have to be struck. Third, this study could have benefitted from supplementary qualitative data, interviews, and self-assessments of participants to further qualify empirical findings.

Fourth, the conceptualization of schadenfreude was merely evaluated on a nominal scale and assessed with only three items. Thus, the accuracy and explanatory power of this variable are limited. As the data suggests, schadenfreude, as a trait of character, plays a noteworthy role in the affective perception of online chats. This relationship could have been explored more extensively since the notion of equal distribution of experimental subjects with a higher schadenfreude score amongst the two conditions of the study was rejected for randomization over a bigger sample size. Fifth, the sample size of *Twitch.tv*-experts was very small. Given a larger cohort, test subjects using a neutral gray interface could have served as a better control group, further strengthening internal validity of the study.

Sixth, in respect to external validity, the extent of generalizability of found effects in the gaming content of *Twitch.tv* on digital chatting could be debated. Disregarding subject matter, large *Twitch.tv* channels feature an extreme chat fluctuation with an

unparalleled rate of messages, which severely limits the potential of truly engaging in dialog. This is further deteriorated by spamming and trolling. Additionally, messages are often dubious in quality of content, information value and significance, which makes it difficult to guarantee similar valence of messages across larger portions of chats. Although the age of the participants was highly appropriate for *Twitch.tv's* target demographic and user-base, the representativeness toward digital chat communication overall remains unexplored. Future work should therefore study the generalizability of the results, for example by trying to reproduce these findings for sentiment scoring of tweets with varying UI background colors.

6 Conclusions and Future Work

The presented findings bear value for application design across the entirety of the digital landscape, balancing issues of aesthetics, legibility and behavioral effects of color on user experience [30]. Depending on intent and use-case of a given service, employing a *black* or a *white* interface respectively might be either beneficial or hindering for the effect a software product is trying to achieve. For example, a grief counseling service comforting depressive or even suicidal patients would be ill-advised to implement a monochrome *black* interface. Ambiguous messages herein would be vulnerable to affective sentimental biases, possibly degrading the patients' condition unintentionally. Dating services, on the other hand, seek to bring people together and deepen mutual sympathy and thus would likely benefit from a *brighter* interface. In instances of critical importance of text and dialog, where objectivity and factual accuracy is paramount, one should refrain from providing purely *white* or *black* interfaces and make use of neutral gray styles instead. A minuscule influence could perchance be a crucial factor in decision making. As it is the case with high stakes business or governmental agreements, considerable repercussions and consequences could potentially ensue. Moreover, it is worthwhile investigating whether the results found in this study are only valid for black and white interfaces or extent to *lighter* and *darker* versions of different hues.

In mobile technology, organic light-emitting diode (OLED) displays are commonly used with Android and Windows Phone devices. In contrast to generic liquid crystal displays (LCDs) that filter light emitted from a built-in backlight, OLED screens display black by deactivating pixel elements altogether. This way, true deep blacks can be achieved without the consumption of any power. Battery life is one of the foremost concerns and bottlenecks in the smartphone industry today [12] since the portable form-factor inherently imposes capacitive limitations. Thus, on OLED mobile devices, black interfaces are heavily favored and deliberately utilized to improve battery efficiency (e.g. Android's power saving mode). These common IT-industry standards and practices might induce, or at least enhance, the vulnerability for affective biases on an immense scale, considering the global user base.

Future work following the direction of this study could explore the presence of these findings in more content-independent means of written and other visual communication. Excluding niche content will likely lead to the discovery of a universal affective brightness bias, in turn resulting in a broader applicability across a wide

variety of digital domains. Adjacent research could investigate the influence on invoked emotion from reading fiction and novels. Especially considering e-readers and their color inverted night mode, the interface might affect intensity and direction of resulting emotion, which can be addressed in future research. In the long run, this might even affect the overall financial success of fiction literature within the confines of digital distribution, excluding print media. Similarly, any artistic endeavor appealing to conjure emotion lends itself to scientific investigation.

By way of example, a digital environment for composing music, or the influence of a gaming interface on moral choice in video games provides promising research ventures. On a grander scheme, automated sentiment analysis could be deployed for an abundance of digital platforms over a large scale of big data to gather comparative sentimental reference, subject to the appearance of different interfaces, across a vast landscape of digital services. Moreover, not only the affective reception of online content biased by different interface colors should be a subject for future research, but also its production and potential interactions. Since digital communication and media assume increasingly dominant relevance in modern lifestyle, the true extent of affective influence and its magnitude gains more importance continuously. Perception and cognitive psychology have an entirely new field of application in the digital domain, as the zeitgeist shifts into intricately interwoven HCI systems across the entirety of the *virtuality continuum*.

References

1. Adams, F.M., Osgood, C.E.: A cross-cultural study of the affective meanings of color. J. Cross Cult. Psychol. **4**(2), 135–156 (1973)
2. Allagui, A., Lemoine, J.: Web interface and consumers' buying intention in e-tailing: results from an online experiment. Adv. Consum. Res. **8**, 24–30 (2006)
3. Baldauf, C.: Metapher und Kognition. Grundlagen einer neuen Theorie der Alltagsmetapher. Peter Lang Verlag, Bern (1997)
4. Bauer, D., Cavonius, C.: Improving the legibility of visual display units through contrast reversal. In: Ergonomic Aspects of Visual Display Terminals, pp. 137–142 (1980)
5. Berlin, B., Kay, P.: Basic Color Terms: Their Universality and Evolution. The David Human Series Philosophy and Cognitive Science Reissues, vol. 19, p. 178 (1969)
6. Bishop, J.: Representations of "trolls" in mass media communication: a review of media-texts and moral panics relating to "internet trolling." Int. J. Web Based Communities **10**(1), 7 (2014)
7. Cambria, E., et al.: Computational intelligence for big social data analysis [guest editorial]. IEEE Comput. Intell. Mag. **11**(3), 8–9 (2016)
8. Cambria, E., et al.: New avenues in knowledge bases for natural language processing. Knowl.-Based Syst. **108**(C), 1–4 (2016)
9. Chang, A.-M., et al.: Evening use of light-emitting eReaders negatively affects sleep, circadian timing, and next-morning alertness. Proc. Nat. Acad. Sci. **112**(4), 201418490 (2014)
10. Cheng, Z.: Effect of font and background color combination on the recognition efficiency for LCD displays. ProQuest Dissertations and Theses, p. 40, May 2015

11. Chiou, W.-B., Cheng, Y.-Y.: In broad daylight, we trust in God! Brightness, the salience of morality, and ethical behavior. J. Environ. Psychol. **36**, 37–42 (2013)
12. Chondro, P., Ruan, S.-J.: Perceptually hue-oriented power-saving scheme with overexposure corrector for AMOLED displays. J. Disp. Technol. **12**(8), 791–800 (2016)
13. Cohen, J.: Statistical Power Analysis for the Behavioural Sciences. Lawrence Earlbaum Associates, Hillside (1988)
14. Desmet, P.: Designing emotion (2002)
15. Desmet, P.M., Hekkert, P.: The basis of product emotions. In: Pleasure With Products: Beyond Usability (2002)
16. Elliot, A.J., Maier, M.A.: Color psychology: effects of perceiving color on psychological functioning in humans. Annu. Rev. Psychol. **65**, 95–120 (2014)
17. Farrington, D.P., Welsh, B.C.: Effects of improved street lighting on crime: a systematic review. Campbell Syst. Rev. **13**, 59 (2008)
18. Faul, F., et al.: G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods **39**(2), 175–191 (2007)
19. Frank, M.G., Gilovich, T.: The dark side of self-perception and social-perception - black uniforms and aggression in professional sports. J. Pers. Soc. Psychol. **54**(1), 74–85 (1988)
20. Garcia-Marques, T., Mackie, D.M.: The positive feeling of familiarity: mood as an information processing regulation mechanism. Psychology Press (2000)
21. Gil, S., Le Bigot, L.: Seeing life through positive-tinted glasses: color-meaning associations. PLoS ONE **9**(8), e104291 (2014)
22. Godbole, N., Srinivasaiah, M.: Large-scale sentiment analysis for news and blogs. In: Conference on Weblogs and Social Media (ICWSM 2007), pp. 219–222 (2007)
23. Greitemeyer, T., et al.: Playing prosocial video games increases empathy and decreases schadenfreude. Emotion **10**(6), 796–802 (2010). Washington, D.C.
24. Hawlitschek, F., et al.: Colors and trust: the influence of user interface design on trust and reciprocity. In: Proceedings of the Annual Hawaii International Conference on System Sciences, pp. 590–599 (2016)
25. Ho, H., et al.: Combining colour and temperature: a blue object is more likely to be judged as warm than a red object. Sci. Rep. **4**, 5527 (2014)
26. Johns, M.: A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep **14**(6), 540–545 (1991)
27. Klein, G.S.: Semantic power measured through the interference of words with color-naming. Am. J. Psychol. **77**(4), 576–588 (1964)
28. Kouloumpis, E., et al.: Twitter sentiment analysis: the good the bad and the OMG! Artif. Intell. **11**(164), 538–541 (2011)
29. Kwallek, N., et al.: Impact of three interior color schemes on worker mood and performance relative to individual environmental sensitivity. Color Res. Appl. **22**(2), 121–132 (1997)
30. Labrecque, L.I., et al.: The marketers' prismatic palette: a review of color research and future directions. Psychol. Mark. **30**(2), 187–202 (2010)
31. Lakens, D., et al.: Brightness differences influence the evaluation of affective pictures. Cogn. Emot. **27**(7), 1225–1246 (2013)
32. Lakens, D., et al.: But for the bad, there would not be good: grounding valence in brightness through shared relational structures. J. Exp. Psychol. Gen. **141**(3), 584–594 (2012)
33. Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press, Chicago (1997)
34. Lakoff, G., Johnson, M.: Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. Basic Books, New York (1999)
35. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)

36. Löffler, D.: Color, Metaphor and Culture. Ph.D. dissertation, University of Würzburg, Würzburg (2017)
37. MacDonald, L.W.: Using color effectively in computer graphics. IEEE Comput. Graphics Appl. **19**(4), 20–35 (1999)
38. Meier, B.P., et al.: Black and white as valence cues: a large-scale replication effort of Meier, Robinson, and Clore (2004, 2015)
39. Meier, B.P., et al.: When "Light" and "Dark" thoughts become light and dark responses: affect biases brightness judgments. Emotion **7**(2), 366–376 (2007)
40. Pelet, J.-É., Papadopoulou, P.: The effect of colors of e-commerce websites on consumer mood, memorization and buying intention. Eur. J. Inf. Syst. **21**(4), 438–467 (2012)
41. Reber, R., et al.: Effects of perceptual fluency on affective judgments. Psychol. Sci. **9**(1), 45–48 (1998)
42. Sánchez, J.A., et al.: Conveying mood and emotion in instant messaging by using a two-dimensional model for affective states. In: Proceedings of VII Brazilian Symposium on Human Factors in Computing Systems, IHC 2006, p. 66. ACM Press, New York (2006)
43. Dos Santos, M., et al.: Method of dynamically evaluating the mood of an instant messaging user (2008)
44. Schwarz, N.: Feelings-as-information theory. In: Van Lange, P.A.M., Kruglanski, A., Higgins, E.T. (eds.) Handbook of Theories of Social Psychology: Collection, vol. 1 and 2, pp. 289–308. Sage, Thousand Oaks (2011)
45. Smith, R.H., et al.: Envy and schadenfreude. Pers. Soc. Psychol. Bull. **22**(2), 158–168 (1996)
46. Stroop, J.R.: Studies of interference in serial verbal reactions. J. Exp. Psychol. **18**(6), 643–662 (1935)
47. Zangerle, E., et al.: SentiStorm: Echtzeit-Stimmungserkennung von Tweets. HMD Praxis der Wirtschaftsinformatik **53**(4), 514–529 (2016)
48. Zarkadi, T., Schnall, S.: "Black and White" thinking: visual contrast polarizes moral judgment. J. Exp. Soc. Psychol. **49**(3), 355–359 (2013)
49. Zhong, C.-B., et al.: Good lamps are the best police: darkness increases dishonesty and self-interested behavior. Psychol. Sci. **21**(3), 311–314 (2010)