Oliver Niebuhr*, Jana Neitsch

Assessing hate-speech perception through bio-signal measurements: A pilot study

Kurzfassung: Aufbauend auf existierenden und realen Hate Speech-Stimuli (geschrieben und gesprochen) zeigen wir in dieser Pilotstudie, dass Biosignale aus Herzrate, Atmung und Hautleitwiderstand mit expliziten Skalenbewertungen der Stimuli aus vorherigen Experimenten konform gehen und somit eine attraktive, direkte Alternative zur Erforschung der Hate Speech-Wahrnehmung darstellen.

Abstract: Based on a set of existing and real hate speech stimuli (written and spoken), our pilot study demonstrates that the bio-signals of heart rate, breathing and skin conductance response mirror explicit scale assessments of the stimuli from previous experiments and are, therefore, an attractive, direct alternative to measuring hate-speech perception

1 Introduction: XPEROHS

Hate Speech is a growing and ubiquitous phenomenon in modern societies around the globe. "As online content continues to grow, so does the spread of hate speech" [1]. The large amount of hate speech presumably results from a toxic mixture of increasing migration, an unbalanced globalization, the supposed anonymity of posts in social networks and the trend towards post-factual statements and promises in politics. However, the means that are currently available to identify, prosecute, and delete hate speech appear underdeveloped. Regarding the growing number of victims of hate speech, not legislation but identification of hate speech is the solution. The young and scattered field of hate speech research analyses the classification and evaluation of hate speech as a "challenging task" [2], mainly due to an insufficient phenomenological understanding. That is, to date "hate speech lacks unique, discriminative features" [2]. Another problem is that there are "differing definitions on what constitutes hate speech" and the limited amount of material to effectively train and use computer based HMM or DNN algorithms in countering hat speech [1].

Our XPEROHS research project (funded by Velux) takes a different approach [3]. Cross-linguistic analyzes and controlled perception studies are combined to uncover hidden hate speech characteristics. It also tests the effect of both written and spoken language on the perception of hate speech. In spoken language, the tone of voice is indeed a further linguistic key element that influences the perception of hate speech [4], but an internal speech melody is also obvious in the written language (e.g. by capitalization). A better understanding of what verbal and non-verbal linguistic features and patterns enhance or tone down hate-speech perception and how listeners from different social groups and languages react to hate speech is supposed to culminate in a more solid, empirically based hate-speech definition. These definitions that are based on our research plus some corresponding recommendations will then be made available to decision-makers in companies and politics for further measures at the end of the project.

Analyzing typical hate speech patterns of German, our perception results suggest that imperatives and Holocaust references enhance hate-speech perception, whereas stylistic devices such as irony and rhetorical questions weaken hate-speech perception. These differences are stronger and more consistent in spoken than in the written mode [4].

2 Questions and assumptions

So far, we have worked with explicit perceiver ratings made by participants clicking onto scales or into innovative 2D rating spaces [4]. This paper uses the same stimulus set as in [4]. However, instead of obtaining explicit ratings, participants' bio-signals are monitored and analysed. In a current pilot study, we aim at analyzing whether bio-signals mirror explicit ratings and are hence a suitable alternative in assessing the perception of hate speech. Accordingly, our research questions are: Are there systematic differences in the perceivers' bio-signals that are exposed to different kinds of hate speech? If so, do these systematic bio-signal differences match (qualitatively) with those of explicit ratings for the same stimuli?

The obvious advantage of bio-signals over explicit ratings are that they can be collected without instruction and active actions by the perceiver, which makes perception studies easier, faster, and less prone to interpretation biases. In addition, bio-signals are a direct manifestation of the (sympathetic) nervous system and therefore less influenced by the

* **Korrespondenzautor: Oliver Niebuhr:** Centre for Industrial Electronics, University of Southern Denmark (SDU), Alsion 2, 6400 Sonderborg, Denmark. E-Mail: oniebuhr@mci.sdu.dk

participants' conscious reflection about correct, appropriate response behavior and/or the supposed goals of the study. This means that bio-signals can provide more reliable and, thus, scientifically more valid and reliable results, as well as more convincing results for the general public.

Three bio-signals were monitored in this pilot study: Heart Rate (HR), Breathing (BR), and Skin-Conductance Response (SCR). HR was determined with the smart watch Garmin Vivoactive HR. The BR measurement was carried out using a two-belt system (Respiratory Inductance Plethysmography), and the SCR data were collected using a two-finger electrode system (Mindfield eSense).

All three bio-signals are known to be positively correlated with mental stress and emotional arousal. So, if bio-signals mirror explicit perceiver ratings on hate-speech severity, then we assume to find an increase in HR/BR/SCR values for hate speech that includes imperatives or Holocaust-references in relation to hate speech containing irony or rhetorical questions. Furthermore, the bio-signal differences between the linguistic feature conditions that are used should be more pronounced for spoken than for written hate-speech stimuli.

3 Method

Twenty participants were recruited for the pilot study. They were subdivided into two groups of 10 people (5 male, 5 female, all between 20-30 years old and German students at the SDU). One subgroup received first the written and then, after a break of several days, the spoken hate speech stimuli. The other subgroup received the two stimulus sets in the inverse order. The spoken stimuli were realizations of the written stimuli, produced by a phonetically trained speaker who met the requirements of a typical hate speaker (i.e. Caucasian white male, between 35-45 years old [5]).

The stimulus sets were presented in blocks of 12 tokens, each of them representing a linguistic-feature condition, i.e., e.g., 12 Holocaust stimuli, 12 irony stimuli, etc. In the spoken set, we included a break of 1 sec in between the 12 tokens of a block. In the written set, the 12 tokens of a block were presented on a single Power-Point slide. The blockwise presentation was to expose participants to a single hate-speech feature condition for about 85-95 sec, ensuring sufficient time to develop clear changes in reaction to the stimuli. Future studies will have to analyse how quickly feature conditions can be changed throughout the study before changes in bio-signals overlap and get blurred.

Participants took part in the study individually in the silent Acoustics Lab at SDU. They were instructed that they their bodily reactions to hate-speech stimuli would be monitored and analyzed. Participants could abort the study at any time. Hate speech signals were monitored time-aligned with each other at a sampling rate of 5 Hz and a 16 bit quantization.

4 Results and Discussion

Our assumptions are met by the findings suggesting that the bio-signals of perceivers mirror the explicit ratings that were obtained with a different group of 28 German participants in [4]. Accordingly, HR/BR/SCR levels rose for hate speech that included imperatives or Holocaust references in relation to hate speech that contained ironic expressions or rhetorical questions with the differences between these linguistic feature conditions being stronger in spoken than in written hate speech. That is, monitoring and analyzing bio-signals is indeed a promising new way of assessing hate speech stimuli on a very direct and subconscious grounds.

Future studies will have to examine the limitations of bio-signal-based hate speech assessments in terms of the dynamics and variability of bio-signal changes and potential advantages regarding the sensitivity of HR, BR, and SCR signals. Further bio-signals (EEG, pupil-dilation) will be included in future studies. We will also investigate to what degree bio-signals could even outperform and supersede explicit perceiver ratings, e.g., with respect to floor or ceiling effects on explicit rating scales and/or in terms of a higher general sensitivity, given that explicit ratings are only able to capture those changes in the assessment and feelings of perceivers that are actually specified by the task or the scale legends.

References

[1] MacAvaney, S., Yao, H. R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, *14*(8).

[2] Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, *10*(5), 925-945.

[3] Baumgarten, N., Bick, E., Geyer, K., Lund Iversen, D., Kleene, A., Lindø, A. V., Neitsch, J., Niebuhr, O., Nielsen, R. & Petersen, E. N. (2019). Towards Balance and Boundaries in Public Discourse: Expressing and Perceiving Online Hate Speech (XPEROHS). *RASK – International journal of language and communication*.

[4] Neitsch, J. & Niebuhr, O. (submitted). *On the role of prosody in the production and evaluation of German hate speech*. Submitted for the 10th International Conference on Speech Prosody, Tokyo, Japan.

[5] Hrdina, M. (2016). Identity, activism and hatred: Hate speech against migrants on Facebook in the Czech Republic in 2015. *Naše společnost*, *14*(1), 38-47.