Stress Lingers: Recognizing the Impact of Task Order on Design of Stress and Emotion Detection Systems

Eduardo Gutierrez Maestro[1], Hadi Banaee[1] and Amy Loutfi[1]

***Abstract*— This paper examines the significance of the priming effect in designing and developing models for recognizing of affective states. Using a public dataset, often considered a benchmark in automatic stress recognition, the significance of the priming effect is explicated. Two experimental setups confirm the importance of task ordering in this problem. The results demonstrate the statistical significance of the model's confusion when the subject has previously experienced stress and illustrate the importance for the Affective Computing community to develop methods to mitigate the priming effect where the order of tasks impacts how data should be modelled.**

I. INTRODUCTION

Affective computing is a multidisciplinary field that seeks the progress of intelligent systems to perceive, process, and simulate human effects. This research field was initially pioneered by Picard *et al.* [1], who investigated different ways to endow machines of such *emotional intelligence*. One step toward achieving this goal is to detect and predict affective states. The problem of recognising affective states has been addressed from different data modalities. In this work, we focus on physiological signals due to their difficulty to mask [2]. Several studies have attempted to predict affective states based on physiological data [3] [4]. A potential limitation of these approaches is that they consider each affective state as an independent event, without accounting for past events.

In psychology, when an individual's exposure to a particular event or stimulus influences their response to subsequent events or stimuli is known as priming. For example, a stressful episode may affect how we react to other stimuli, for example, an amusement video (see Fig. 1). To the best of our knowledge, the priming effect has not been considered in the field of affective computing.

We investigate the impact of the sequential order of affective reactions when developing predictive models. We make use of the public benchmark WESAD [3], which is composed of stressful and amusing tasks in different sequential orders. Our hypothesis posits that the temporality of the stressful task impacts the confusion of the amusement task. We examine this using two experimental setups: leave-one-subject-out (LOSO) and train-test random split.

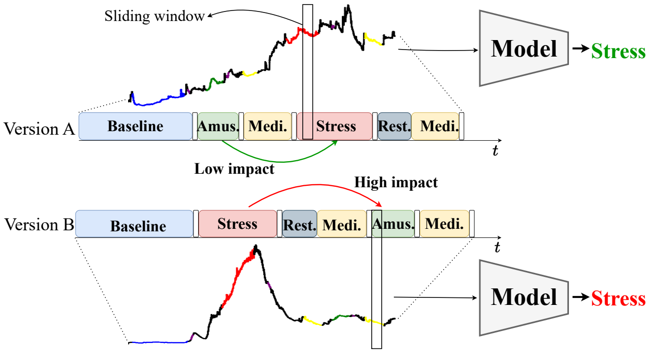

Fig. 1: Graphical illustration of priming effect on two versions of tasks order. Emphasizing *version B*, the stressful task has an impact on subsequent events.

The main contribution of this work is emphasizing the significance of accounting for the temporal ordering sequence of tasks. Accounting for this priming effect in affective predictive models may enhance machines perceiving and interpreting affective reactions.

II. METHODOLOGY

A. Dataset description

We use the public benchmark *WESAD* [3]. The data collection design of this dataset is well-suited for our research goal, as it incorporates different variations in the sequential order of tasks that elicit distinct physiological reactions.

This dataset was collected with the goal of registering physiological reactions corresponding to stress and amusement states. A total of 15 participants participated in the construction of this dataset. Authors in [3] collected signals in a controlled environment, i.e. in the lab. They used two different sensors: *RespiBAN Professional*, a chest-worn sensor; and *Empatica E4*, a wrist-worn sensor.

The data collection protocol consisted of performing two tasks. The stressful task consisted of the Trier Social Stress Test [5], which involves public speaking and arithmetical exercises. The amusement task consisted of watching a set of funny videos. In [3], authors highlighted two versions of the data collection protocol. The two versions follow different task orders (see Fig. 1). In version A, participants were amused at first. On the contrary, in version B participants performed the stressful task first, as illustrated in Fig. 1.

B. Experimental setup

As mentioned previously, we used signals collected from the chest-worn sensor. It is used a sliding window of 60

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation

[1] Centre for Applied Autonomous Sensor Systems (AASS), Örebro University, Sweden {eduardo.gutierrez-maestro, hadi.banaee, amy.loutfi}@oru.se

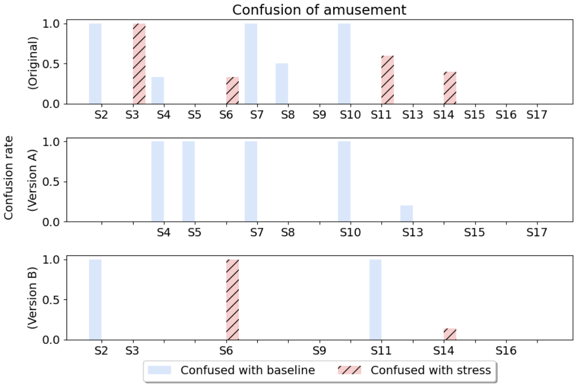

Fig. 2: Confusion rate of the *amusement* class.

seconds (as reported in previous works [3]) to extract samples, that we will use to train our model. Each sample is labelled with the task associated with the temporal position of the sliding window. We study our hypothesis by training a CNN-based model inspired by [4].

III. Experimental Results

This work aims to show the importance of accounting for the priming effect when designing affective predictive models. To that end, we design a set of experiments showing the existence of such an effect in one public benchmark used in the field [3]. In concrete terms, we observed a significant level of confusion in the state of amusement whenever the subject had previously experienced high levels of stress during the previous task. We use three different subsets of the original dataset for our experiments: the original dataset, and the two versions mentioned previously (version A and B in Fig. 1. For each experiment, we illustrate the confusion rates for each actual class (stress or amusement) for better comparison. For a better understanding of the figures, we leave blank the position of a participant that is not part of the dataset, e.g. subject $S2$ is skipped in version A plots. Two types of experiments are conducted. The first one is subject-oriented, in which the model's validation method has been used in previous works [4]. The second experiment follows a different validation approach, by randomly splitting samples among subjects.

Subject-dependent Classification. In this experiment we conducted a LOSO validation method. It consists of testing the model on one subject and training it on the remaining subjects. Initially, we train a model on the original dataset without considering the differences between version A and version B as it has been done in previous works [3][4]. On average, we obtained an f1-score of 0.81 ± 0.25. This metric is the average among all tested subjects. Despite the overall good performance, similar to the one obtained in previous works, we have gone further investigating the confusion ratios for the *amusement* state. We illustrate such confusion rates for each subset mentioned previously in Fig. 2. It is observed that the highest confusion ratios for the amusement class correspond to subjects belonging to version B (first-row Fig. 2). Since our goal is to show the importance of accounting for the temporal order of tasks, also known as priming, we repeated the experiment but now separating subjects into version A and B data subsets. Now, the average f1-score is 0.84 ± 0.13 and 0.68 ± 0.30 for versions A and B respectively. The version A subset exhibits no confusion on the amusement class by stress (middle-row of Fig. 2), while the version B subset still shows signs of confusion (last-row of Fig. 2). The difference in performance between these two recent experiments provides a positive indication of our hypothesis. We conducted a statistical *t-test* to evaluate whether the confusion rates of the *amusement* class with the *stress* class obtained are statistically significant. We performed the test over the three datasets mentioned previously. We conducted 10 runs of the model and collected the confusion rate of the *amusement* class by the *stress* class to perform the *t-test*. The tests showed evidence ($p - value < 0.005$) of statistical difference between the original dataset and the version A set, and between version A and B sets. These results confirm our hypothesis: priming makes stress linger impacting how data for the amusement class is generated.

Subject-independent Classification. The previous experiment was challenging due to the physiological variability between train and test sets. In [1], the authors argued that such variability hinders the generalization of the models. Since our goal is to study the impact of task order and not the generalization, we designed this experiment whose train and test sets (randomly separated 75% and 25% respectively) were less challenging. The average f1-score reported was in this case 0.83 ± 0.26 for the original dataset, 0.95 ± 0.03 for version A and 0.97 ± 0.04 for version B. We also conducted statistical tests to verify significant differences. These tests yielded the same evidence as in the previous experiment.

IV. Conclusion

To the best of our knowledge, the priming effect has yet to be thoroughly been examined in this community. This paper makes an important step in demonstrating the importance of accounting for temporal task order when designing affective predictive models. In future work, this study will be extended along with other benchmarks by proposing a solution to mitigate the priming effect.

References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, 2020.

[3] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018.

[4] R. K. Sah and H. Ghasemzadeh, "Stress classification and personalization: Getting the most out of the least," *IEEE 17th Int'l. Conf. BSN*, 2021.

[5] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, 1993.