

LightGBM Prediction of Digital Self-Harm from Self-Criticism, Emotion Dysregulation, and Online Disinhibition

Deborah. Cohen¹, Rachel. Higgins², Ariel. Paredes de Jesús³, Javier. Galvez-Martos^{3*}

¹ Department of Psychology, Queen's University, Kingston, Canada

² Department of Counseling, Leadership and Research Methods, University of Arkansas, Graduate Education Building, Fayetteville, AR 72701, USA

³ Department of Psychology, Universidad de La Frontera, Temuco 4780000, Chile

* Corresponding author email address: javier.galvez@ufrontera.cl

E d i t o r	R e v i e w e r s
José Aparecido Da Silva  Full Professor, Department of Psychology, University of Sao Paulo, Ribeirao Preto, Sao Paulo, Brazil jadsilva@ffclrp.usp.br	Reviewer 1: Bridget Abalorio  Faculty of Psychology, Peruvian University of Applied Sciences, Lima, Peru. Email: bridget.abaloro@upc.edu.pe Reviewer 2: Ana Amelia Coder  Department of Psychology, University of New Brunswick, Saint John, New Brunswick, Canada. Email: anaameliacoder@gmail.com

1. Round 1

1.1. Reviewer 1

Reviewer:

In the Introduction paragraph discussing conventional cyberbullying, the manuscript notes that “digital self-harm disrupts the usual perpetrator-victim distinction by placing self-directed aggression within the architecture of online communication.” This is one of the strongest conceptual sentences in the paper, but it should be developed further. The authors should explain whether digital self-harm is expected to share mechanisms with cyberbullying perpetration, cybervictimization, self-injury, or all three. Without this clarification, the literature review risks relying too heavily on studies of other-directed cyber aggression, while the outcome variable is self-directed. A short theoretical bridge should be added to justify why evidence from cyberbullying and online aggression is applicable to digital self-harm.

In the paragraph on online disinhibition, the manuscript appropriately distinguishes between benign and toxic forms of disinhibition. However, the sentence “toxic online disinhibition may be especially relevant because it can lower the psychological barrier against posting hostile or humiliating statements” should specify whether the hostility is directed toward others, toward the self, or both. Since the findings show that toxic online disinhibition was the strongest predictor, the

Introduction should more clearly hypothesize why a construct usually associated with interpersonal online aggression would predict self-directed online harm. This would improve alignment between the theoretical framework and the empirical results.

In the Data Collection Tools section, the manuscript indicates that the digital self-harm score was treated as a continuous severity variable. This decision should be justified more explicitly. Digital self-harm data are often skewed, with many participants reporting zero or low involvement. The authors should explain whether the score distribution supported continuous regression modeling, whether zero-inflation was examined, and whether sensitivity analyses were conducted using binary or ordinal classification. Without this information, the choice of regression-based LightGBM may be questioned, particularly if the outcome has a highly non-normal or sparse distribution.

In the Methods section, the description of translated and adapted tools states that the instruments were reviewed for “linguistic clarity and cultural fit in Chilean Spanish.” This needs greater methodological specificity. The authors should report whether forward translation and back-translation were conducted by independent bilingual experts, whether a committee review was used, whether cognitive interviews were performed, and whether measurement equivalence was considered. Since the study is conducted in Chile and uses constructs developed largely in other linguistic and cultural contexts, the validity of Chilean Spanish adaptation is a central methodological issue, not a minor procedural detail.

Authors revised and uploaded the document.

1.2. Reviewer 2

Reviewer:

In the paragraph on self-criticism, the manuscript states that “self-criticism is another central construct in understanding digital self-harm.” This claim is plausible, but the argument remains underdeveloped. The authors should define self-criticism in terms of inadequate self, hated self, and reassured self before the Methods section, because these three dimensions later become central predictors. In particular, hated self should be described as conceptually closer to self-directed hostility and punitive self-attack, which would help explain why it emerged as a stronger predictor than inadequate self in the SHAP analysis.

In the paragraph on emotion dysregulation, the manuscript explains that difficulties may include nonacceptance, impulse control difficulties, limited strategies, lack of awareness, and lack of clarity. This is appropriate, but the paragraph should explicitly state which dimensions are expected to be most relevant to digital self-harm and why. The findings later show that limited access to regulation strategies, impulse control difficulties, and nonacceptance of emotional responses were among the strongest predictors. This pattern would be more convincing if the Introduction had pre-specified a theoretical rationale that digital self-harm may function as an immediate maladaptive strategy during high distress, especially when adaptive regulation strategies are perceived as unavailable.

In the final paragraph of the Introduction, the manuscript states that “Light Gradient Boosting Machine modeling is especially useful for tabular psychological data because it can identify high-contribution predictors while preserving strong predictive performance.” This is an important methodological justification, but it requires a more careful scientific explanation. The authors should briefly explain why LightGBM was selected over alternative machine learning models, such as random forest, XGBoost, support vector regression, or elastic net regression. The current justification is acceptable but generic; it should be revised to include the model’s ability to capture nonlinear associations, interaction effects, and variable importance patterns in moderate-sized psychological datasets.

In the Methods section, the paragraph on Study Design and Participants states that “the final sample included 512 participants from Chile.” The exact sample size is clearly reported, but the sampling procedure remains insufficiently detailed. The authors should specify whether recruitment was convenience-based, stratified, snowball-based, or conducted through institutional channels. In addition, the manuscript should report how many individuals initially accessed the survey, how many were excluded, and the reasons for exclusion. This information is necessary for evaluating possible selection bias and the credibility of the final analytic sample.

In the participant paragraph, the manuscript reports recruitment from “universities, technical institutes, community online groups, and public digital advertisements in Santiago, Valparaíso, Concepción, La Serena, and Temuco.” This geographical coverage is valuable, but the authors should explain whether the sample was proportionally distributed across these cities or dominated by one urban center. Because online behavior and mental health service access may differ across regions, socioeconomic backgrounds, and institutional contexts in Chile, a clearer description of regional and educational composition is necessary. At minimum, the manuscript should add demographic frequencies for region, gender, educational status, and daily internet use in the Findings section.

In the Methods section, the description of the Digital Self-Harm Behavior Checklist states that the checklist was “adapted from the operational definition proposed by Patchin and Hinduja.” This is acceptable, but the instrument description requires stronger psychometric detail. The authors should report the exact number of items, provide one or two sample item themes without using overly sensitive wording, explain the scoring range, and state whether the checklist has been previously validated or whether it was newly adapted for this study. If newly adapted, the authors should describe content validity procedures, expert panel composition, pilot sample size, and whether any items were modified after pilot testing.

Authors revised and uploaded the document.

2. Revised

Editor’s decision after revisions: Accepted.

Editor in Chief’s decision: Accepted.