The Journal of Clinical Psychiatry

Supplementary Material

Article Title: Social Media Images Can Predict Suicide Risk Using Interpretable Large Language-Vision Models

Author(s): Yael Badian, MSc; Yaakov Ophir, PhD; Refael Tikochinski, MA; Nitay Calderon, MSc; Anat Brunstein Klomek, PhD; Eyal Fruchter, MD; and Roi Reichart, PhD

DOI Number: https://doi.org/10.4088/JCP.23m14962

LIST OF SUPPLEMENTARY MATERIAL FOR THE ARTICLE

1. **Appendix 1** Supplementary Methods and Results

DISCLAIMER

This Supplementary Material has been provided by the author(s) as an enhancement to the published article. It has been approved by peer review; however, it has undergone neither editing nor formatting by in-house editorial staff. The material is presented in the manner supplied by the author.

Social Media Images Can Predict Suicide Risk Using Interpretable Large Language-Vision Models

Supplementary Appendix 1

The suicide risk outcome measure of the study

The participants' risk of suicide was assessed with the well-established and well-researched CSSRS – the Columbia Suicide Severity Rating Scale.[1] The CSSRS has high predictive validity of suicide risk[2,3] and it consists of 6 categorical (yes/no) items. The first two items measure the very existence of a suicide risk, that is the risk that the person is experiencing any level of suicidal thoughts, whether these thoughts are concrete and highly dangerous, or 'just' passive and abstract death wishes. The remaining four items measure the severity of this general risk, and they are shown to the respondents only if the first two items indicated that they are at a (general) suicide risk. These items address concrete ideation to engage in active suicide behaviors, such as when the person reports of having a specific method or a plan to act on their suicidal thoughts. Notably, a positive answer to one or more of these four items indicates that the person is at a relatively high risk of suicide. In the current study, we therefore used this stricter cut-off point for a *high suicide risk* as our primary outcome (to be predicted by Facebook images).

The sample of the study

Of the initial sample of 2,685 MTurk users, 462 participants did not provide a working Facebook ID, 102 participants did not upload images to their timeline, and 341 participants failed implanted quality checks we developed to detect inattentive and bogus crowdsourcing respondents.[4] We also removed users who uploaded a relatively small number of images to their Facebook account (i.e., users who had less than 39 images – the median number of images in the sample) to ensure that our further computational analyses will be based on a substantial amount of visual data for each participant. The final and cleansed dataset included 841 high-quality respondents (83.4% female, average age = 36.7) who uploaded together 177,220 images (M =

124, *SD* = 218.8). Corresponding with previous studies that documented increased levels of mental health issues on MTurk (e.g., [4-6]), relatively high proportions of the current sample were classified as 'participants at high suicide risk' (10.93%).

Table 1 (in the body of the article) provides the descriptive statistics of the dataset. Complimentary description and statistical analyses of the entire sample of high-quality respondents who uploaded at least one accessible image (*N* = 1697) are available by the authors upon request. For further, detailed information about the complete dataset, see in Ophir, Tikochinski, et al., 2020.[7]

The vison-language model of CLIP

The Facebook images were represented using the recently developed deep learning model of CLIP (Contrastive Language-Image Pre-training; see the Supplementary Material).[8] CLIP is a multi-modality deep neural network consisting of two components (encoders) that can represent images and texts as dense-numeric vectors. CLIP was trained in a bottom-up manner, to match the right textual captions with their corresponding images using tens of randomly sampled options. The developers of CLIP collected 400 million pairs of images and texts from various sources on the internet. To ensure a large variety of visual concepts, each text sample had to include one word from a set of 500K queries. This set of queries consisted of all the words that occurred at least 100 times in the English version of Wikipedia.

In practice, CLIP uses representation vectors to evaluate the similarity between images and texts and assigns probabilities to each candidate caption based on its similarity to the image. It then selects the caption that achieved the highest probability score as the correct caption of the given image. This training allows CLIP to be used for various sub-tasks, such as extracting visual features from an image. For example, to detect whether an image is bright or dark, researchers can provide CLIP with the image and a set of captions (queries) – "a bright image" and "a dark image". CLIP then assigns probabilities to each one of the queries (e.g., "a bright image" = 0.7 and "a dark image" = 0.3). Based on these probabilities, which sum up to 1, the

researchers can determine which one of the queries is most likely to be correct for this image ("a bright image").

Supplemented information about the extraction of the interpretable visual features

As explained in the main article, this study used CLIP in an unconventional way as a preliminary methodological step to extract visual features, which were predefined by us, in advance, in a top-down manner (for details about the extraction process, see the Method section). This is in contrast to common uses of CLIP, as CLIP is typically utilized for solving end-to-end tasks, such as object detection or segmentation.[9]

It should be noted here that the exact verbal phrasing of the queries affects the probability scores generated by CLIP. For example, the score of the query "a bright image" can differ from the score of "the image is bright". To ensure that the chosen queries of the current study were well phrased, we randomly selected 10 images from the Internet (i.e., not from our dataset, see examples in Table 2), and applied CLIP to test different phrasing alternatives until we received satisfactory results. We conducted this fine-tuning phrasing of the queries on external images from the Internet to overcome the potential problem of overfitting, which might have occurred if we were to conduct it on original images from our dataset.

Supplemented information about the extraction of the theory-driven features

Aside from the key theory-driven features discussed in the main article, we reviewed published lists of risk factors of suicide by leading health establishments, such as the World Health Organization (WHO), the Centers for Disease Control and Prevention (CDC), and the National Institute of Mental Health (NIMH),[10] and searched for additional risks that might be evident in social media images. For each risk factor (e.g., prior suicide attempts, drug abuse, and psychiatric diagnoses), we phrased matching visual queries (e.g., the person in the image abuses drugs), but CLIP could not perform well with these queries (nor could we, as human experts), probably because the images did not contain such blatant risk factors. The only theory-driven features we could extract from these social media images targeted emotions and relationships, as hypothesized by the interpersonal-psychological theory of suicidal behavior,[11,12] as well as the

evidence-based treatments – the interpersonal psychotherapy (IPT)[13] and the attachment-based family therapy (ABFT).[14,15] Further studies that will find ways to consider stronger theory-driven risks as potential predictors are therefore encouraged, as they might achieve even better results than the obtained prediction scores of this study.

The prediction performance measure (AUC scores)

To evaluate the prediction performance of the various models of the study, we used the standard measure of AUC – the Area Under the Receiver Operating Characteristic (ROC) curve. The AUC measure is most appropriate for such a class-imbalanced dataset,[16] since it provides a single holistic value that reflects the relations between correct predictions of suicide (true-positive) and incorrect predictions of suicide (false-positive) at all potential classification thresholds.

Supplemented information about the t-test comparisons of the visual features

As mentioned in the Results section, a further analysis was conducted to explore the associations between the 24 CLIP-based features and the risk of suicide in the entire sample. The first step of this analysis included a t-test comparison of the mean probability scores of the visual features between the high suicide risk group (N = 92) and the rest of the sample (N = 749), using an FDR correction for multiple tests.[17] This procedure yielded 17 significant features. However, 6 of these 17 features had to be removed from the final table of differences (Table 4 of the main article) because 6 CLIP tasks involved only two (opposite) queries that sum up to the probability of one, thus creating redundant duplicates (e.g., the t scores of ‘happy people’ were the same as of ‘sad people’).

References

1. Posner K, Brown GK, Stanley B, et al. The Columbia–Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry*. 2011;168(12):1266-1277.

2. Drapeau CW, Nadorff MR, McCall WV, Titus CE, Barclay N, Payne A. Screening for suicide risk in adult sleep patients. *Sleep Medicine Reviews*. 2019/08/01/ 2019;46:17-26. doi:https://doi.org/10.1016/j.smrv.2019.03.009

3. Weber AN, Michail M, Thompson A, Fiedorowicz JG. Psychiatric emergencies: assessing and managing suicidal ideation. *Medical Clinics*. 2017;101(3):553-571.

4. Ophir Y, Sisso I, Asterhan CSC, Tikochinski R, Reichart R. The turker blues: Hidden factors behind increased depression rates among Amazon's Mechanical Turkers. *Clinical Psychological Science*. 2020;8(1):65-83.

5. Arditte KA, Çek D, Shaw AM, Timpano KR. The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment*. 2016;28(6):684.

6. McCredie MN, Morey LC. Who Are the Turkers? A Characterization of MTurk Workers Using the Personality Assessment Inventory. *Assessment*. 2018:1073191118760709.

7. Ophir Y, Tikochinski R, Asterhan CSC, Sisso I, Reichart R. Deep neural networks detect suicide risk from textual facebook posts. *Scientific Reports*. 2020/10/07 2020;10(1):16685. doi:10.1038/s41598-020-73917-0

8. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. PMLR; 2021:8748-8763.

9. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. *ACM computing surveys (CSUR)*. 2022;54(10s):1-41.

10. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychological bulletin*. 2017;143(2):187.

11. Ribeiro JD, Joiner TE. The interpersonal-psychological theory of suicidal behavior: Current status and future directions. *Journal of clinical psychology*. 2009;65(12):1291-1299.

12. Joiner T. *Why people die by suicide*. Harvard University Press; 2009.

13. Markowitz JC, Weissman MM. Interpersonal psychotherapy: principles and applications. *World Psychiatry*. 2004;3(3):136.

14. Ewing ESK, Diamond G, Levy S. Attachment-based family therapy for depressed and suicidal adolescents: theory, clinical model and empirical support. *Attachment & human development*. 2015;17(2):136-156.

15. Diamond G, Diamond GM, Levy S. Attachment-based family therapy: Theory, clinical model, outcomes, and process research. *Journal of Affective Disorders*. 2021/11/01/ 2021;294:286-295. doi:https://doi.org/10.1016/j.jad.2021.07.005

16. Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data--recommendations for the use of performance metrics. IEEE; 2013:245-251.

17. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*. 1995;57(1):289-300.