**Using Deep Neural Networks for Suicide Risk Detection from Textual Facebook Posts**

\*Yaakov Ophir, PhD1,2

Refael Tikochinski1

Christa Asterhan, PhD1

Itay Sisso1

Roi Reichart, PhD2

1The Hebrew University of Jerusalem, 2Technion—Israel Institute of Technology

**Acknowledgments:** The research presented here was conducted with the financial support of the Israeli Innovation Authority (Kamin grants #60561 and #60560).

**Declaration of Interest:** None

**Data Availability**: The data that support the findings of this study are available on request from the corresponding author, [YO]. The data are not publicly available due to their containing information that could compromise the privacy of research participants.

Correspondence should be addressed to Yaakov Ophir, The Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa, Israel. E-mail: yaakov.ophir@mail.huji.ac.il

**Abstract**

*Background:* An emerging field of research suggests that the vast amount of real-time data accumulating on social media contain valuable information regarding the psychiatric and emotional functioning of the user. *Aims*: The goal of this study is to construct deep-learning Artificial Neural Network (ANN) models, which could predict suicide risk from social media textual postings. *Method*: The dataset included 85,643 posts matched with clinically valid psychosocial information about 1,650 authenticated Facebook users. *Results*: Using Deep Contextualized Word Embeddings for text representation, two models were constructed: A Single Task Model (Facebook texts → suicide) confirmed that suicide can be predicted from Facebook postings, especially among users who are relatively active on this platform. A Multi-Task Model which included hierarchical, multilayered sets of theory-driven risk factors (Facebook texts → personality traits → psychosocial risks → psychiatric disorders → suicide) improved predictions (.690 ≤ AUC ≤ .759), with substantially larger effect sizes (.701 ≤ *d* ≤ .994). Subsequent content analyses suggest that predictions did not rely on explicit suicide-related themes, but on a wide range of textual content. *Conclusions*: The methodological and theoretical contributions of this study to suicide risk research and to practical development of automated suicide screening tools are discussed.

**Keywords**: Suicide; Depression, Artificial Neural Network; Machine Learning, Social Networking Sites

**Introduction**

Early detection of suicide risk is a prerequisite for improving suicide prevention efforts (1, 2). The goal of this study is to leverage advancements in deep-learning techniques to predict suicide risk from social media postings. Recent findings show that social media behaviors contain valuable information regarding the mental health (3, 4, 5) and depressive symptoms (6, 7, 8) of users. However, only few studies aimed to predict suicide risk from social media and their prediction validity is limited.

The existing studies on suicide rarely include offline, external validations of suicide risk. They rely on *proxy diagnostic signals*: “tweets” or posts containing explicit references to suicide (9, 10, 11), usually from designated online support forums, such as Reddit (12, 13). These *proxy signals* suffer from poor external validly (14) because they are not always indicative of actual risk, especially when they appear in platforms not related directly to suicide (a Facebook post such as “OMG, I just want to kill myself” does not necessarily indicate concrete suicidal thoughts). Moreover, many users choose not to share their personal distress online, and without external measures of suicide they cannot be detected. Finally, research on digital footprints of psychopathologies in general and suicide in particular rarely considers the broader clinical picture of the predicted phenomenon. Without considering the wide-ranging potential risk factors for suicide and without external validation of actual risk, the construct and external validity of the studies are limited.

This research aims to construct a deep-learning neural network that could predict suicide risk from social media texts, while considering the aforementioned limitations in the literature. A total of 1,650 Facebook users completed a well-established, clinically valid screening tool of suicide risk (15) and volunteered to disclose a year of their Facebook activity, resulting in a dataset of 85,643 Facebook postings. Clinically validated data was also collected on three sets of risk factors for suicide and for depressive episodes, which often precede suicidal behavior (16).

The first set comprised psychiatric disorders, the most severe risk factors for suicide behaviors (9). This set included depression alongside generalized anxiety, which often appears in comorbidity with depression (16, 17). The second set included *psychosocial risks* for depression (18, 19), namely: depressive rumination, excessive worries (19, 20), feelings of loneliness, and lack of satisfaction with life (21, 22). The third and most distal set of factors included the Big Five personality traits (23), since Neuroticism and, to a lesser extent, Extroversion have been associated with suicide behaviors (24) and depressive symptoms (16).

Based on this dataset, we extracted representations of Facebook texts, using a deep contextualized word embedding model (see Method section) and constructed Artificial Neural Network (ANN) models to predict suicide risk from these representations. Our first hypothesis (H1) was that a straightforward Single Task Model (STM) would predict suicide risk from users’ Facebook activity (Facebook texts → suicide). Our second hypothesis (H2) was that a Multi Task Model (MTM) that considers multiple, theory-driven layers of contributing factors (Facebook texts → personality traits → psychosocial risks → psychiatric disorders → suicide) would yield improved suicide risk predictions, compared with the previous STM. Finally, we provide interpretational analyses of the predictions of the computational models to identify textual features that may have contributed to the distinction between individuals with and without suicide risk.

**Method**

***Tools and Measurements***

**Facebook data collection.** Facebook users (*N* = 1,650) who agreed to participate in the study gave us a one-time authorization to download their Facebook posts up to 12 months prior to the date of agreement. A total of 85,643 original postings generated and posted on their timeline by the participants themselves were extracted through a designated application. The *median* number of Facebook postings per profile was 10 (*M* = 42.99, *SD* = 86.28). The *median* number of words in each post was 27 (*M* = 35.23, *SD* = 38.42). 1,002 participants published at least 10 posts and were marked as “*Active Facebook users*.”

**Suicide risk.** Suicide risk was measured using the 6-item Columbia Suicide Severity Rating Scale (CSSRS) (15). The CSSRS is considered a diagnostic tool of choice in clinical settings and empirical research, with high specificity and sensitivity (25, 26). The modular structure of the scale enables extraction of two binary (yes/no) variables: a *general risk of suicide* (participants who met the criterion of any suicidal thoughts) and a *high risk of suicide* (a sub-group of the 'general risk' participants who reported a specific method, intentions, or plan to act on their suicidal thoughts). The sum score correlated positively with all the examined risk factors and especially with depression (*r* = 0.46), thus indicating a high convergent validity of the scale (see Table A, Supplementary Information).

**Risk factors for suicide and depression.** Major depressive disorder was measured using the Patient Health Questionnaire-9 (PHQ-9) (27). Generalized anxiety disorder was measured using the GAD-7 (28). Depressive rumination (brooding) was measured using five items from the Ruminative Responses Scale (RSS) (29). Excessive worrying was measured using the Penn State Worry Questionnaire (PSWQ) (30). Loneliness was measured using the 10-item version of the UCLA-Loneliness Scale (31). Low satisfaction with life was measured using the Satisfaction With Life Scale (SWLS) (32). Personality traits were assessed using the short version of the Big Five Inventory (BFI-10) (33). Complete descriptions of scales used in this study and their convergent validity scores are provided in the Supplementary Information.

***Sample and Dataset***

The procedures of the study comply with the ethical standards of the national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures were approved by the Ethics for Research on Human Subjects Committees at the Technion Israel Institute of Technology and the Hebrew University of Jerusalem. Participant recruitment was conducted through Amazon’s Mechanical Turk (MTurk). A strict data quality assurance protocol for online data collection was applied (34), which included a method to screen out bogus participants (35) and eight attention checks (see Supplementary Information). The recruited participants read and signed a detailed consent form, completed surveys on eight psycho-diagnostic measures, and installed the application that extracted their Facebook content to an external, encrypted data storage space. Upon completion, participants who met the criterion for suicide risk received a letter that included a list of mental health services and an encouragement to seek help (see Supplementary Information).

A total of 2,685 adult MTurk users (36% female, average age = 34.80 yrs) completed the full survey, of which 236 users had suspicious IP addresses. From the remaining users, 1,985 passed the eight attention checks. The 335 users who did not publish any Facebook postings were omitted from this study. The final sample included 1,650 attentive users who published at least one Facebook post.

Descriptive statistics and zero-order correlations of the psycho-diagnostic measures are provided in the Supplementary Information (Table A). Based on previous works and the psychological compositions of MTurk samples, we note that the prevalence of mental health issues, especially major depression, is significantly higher in MTurk, compared with the general population (34, 36, 37). Correspondingly, relatively high rates of suicide risk were found in the current sample: 568 users (34.4%) met the criterion ‘*general risk of suicide*,’ of which 204 (12.4%) met the criterion for *high risk of suicide*. Similar percentages (36.03% and 13.17%, for general and high risk, respectively) were observed among the sub-set of *Active Facebook Users* (*N* = 1,002). The difference in suicide risk rates between active and non-active Facebook users was not significant, *t*(1648) = 1.705, *p* > .05 and *t*(1648) = 1.243, *p* >.05, for general and high risk, respectively.

**ANN-based Models**

Two ANN-based models were constructed (Figures 1 and 2). The architectures of both models are described in the Supplementary Information. Both models consisted of identical input and output layers. The input consisted of representations of Facebook texts, which are 1024-dimensional vectors extracted by the ELMo contextualized word embeddings model (38), a state-of-the-art ANN framework especially relevant to social media language (see theDiscussionand Supplementary Information). The output consisted of a single binary (yes/no) variable of suicide risk. Following the modular structure of the suicide scale, we considered two variants of each model, one for predicting *general risk of suicide* and one for predicting *high risk of suicide*.

The two variants of the Single Task Model (STM) were constructed to predict suicide risk directly from textual contents of Facebook posts only (textual content → suicide). The two Multi-Task Model (MTM) variants were constructed to predict a hierarchical combination of multiple factors. We integrated three sets of risk factors that could mediate the link between Facebook postings and suicide risk (textual content → personality traits → psychosocial risks → psychiatric disorders → suicide). Illustration of this model is provided in the Supplementary Information (Figure A).

In the learning phase, each ANN-based model was trained on 70% of the input data (Facebook texts of 1,155 users), to distinguish between Facebook patterns of suicidal and non-suicidal individuals. Each learning example is comprised of the Facebook texts of one participant together with the suicide label of that participant (general/high suicide risk). For the MTM model, it includes the auxiliary variables scores of this participant (their scores on the psychosocial scales).

In the development phase, a hyper-parameter tuning process was conducted on another 15% of the data (247 users). In this phase, we also considered several alternative models that were more complicated than the STM but less complicated than the MTM. These partial models included one of the MTM three auxiliary layers (e.g., psychiatric disorders) but their detection performance did not reach the quality of the complete MTM. In the test phase, the remaining 15% of the dataset (248 users) was used to examine the predictive quality of each model. The full details of the model, including its objective function, training algorithm, hyper-parameters, and tuning procedure are provided in the Supplementary Information.

**Results**

***Detection Performance of Suicide Risk***

A Receiver Operating Characteristic curve (ROC curve), which plots the True Positive prediction rates of the models against the False Positive rates was generated and the Area Under the ROC Curve (AUC) was calculated. AUC provides a reliable estimation of the quality of the predictions across all possible classification thresholds. It specifically suits class imbalanced tasks in which the positive class (suicidal users) is significantly smaller than the negative class (non-suicidal users) (39). It can be transformed to the common effect-size measure (*Cohen’s d*) in experimental psychology (40).

Table 1 demonstrates the detection performance of the two models. In support of our first hypothesis (H1), the performance of the STM shows that Facebook texts indeed include discernable signals that can be used for predicting suicide risk, even when the model is applied to all users, regardless of their activity level (AUC = .567 and .555, for general and high suicide risk, respectively). Performance measures improve when the model is applied to Active Facebook Users only (AUC = .608 and .606 for general and high suicide risk, respectively). A transformation of these AUC scores to effect sizes (40) indicated a small to medium effect size for general risk (Cohen’s *d* = .388) and high risk (Cohen’s *d* = .380) of suicide.

The inclusion of all risk factors in one MTM yielded improved predictions, especially among Active Facebook Users (*AUC* = .759 and .690, for general and high suicide risk, respectively). These predictions show a medium-to-large effect for high risk of suicide (Cohen’s *d* = .701) and a large to very large effect for general risk of suicide (Cohen’s *d* = .994). A similar pattern of results was found when we represented the Facebook texts with the recent attention-based BERT model (Bidirectional Encoder Representations from Transformers) (41), indicating that the observed patterns and predictions extend beyond the specific representation method (ELMo) that was employed in this study (see Supplementary Information, Table B). The results support our second hypothesis (H2) that a multilayered prediction model consisting of all three layers of contributing factors (Facebook content → personality traits → psychosocial risks → psychiatric disorders → suicide) would demonstrate improved predictions, in comparison with a STM.

***Interpretation of Observed Predictions***

Based on the threshold that best distinguished users at general suicide risk[[1]](#footnote-1) from the rest of the sample, we categorized the users to four possible prediction classifications: True Positive, False Positive, True Negative, and False Negative (see Supplementary Information). Then, we conducted a word search for suicide-related content among active users at general risk who were classified correctly by the MTM (*N* = 33 True Positive users, 22% of the test data). This search produced eight mentions of *suicide/ suicidal*, 20 mentions of *kill*, and 44 appearances of *die* (including *dying, dead*, and *death*). Only in a single instance did these words appear in messages directly related to suicide. Two examples are “*my back is killing me*” and “*It’s gonna be a good Halloween, probably going to die, but it’ll be fun*.” Even in the case of the most explicit phrase “I want to die,” the full context was: “*Cramps so bad, I want to die*”.

Finally, we applied *Term Frequency Inverse Document Frequency (TF-IDF)* analysis (42) to extract the hundred most frequent words that best distinguished between the four possible classes of prediction of general suicide risk: True Positive, False Positive, True Negative, and False Negative among active users (see Supplementary Information, Table C, for the full list). Users at general suicide risk who were identified correctly by the MTM (True Positive) had high frequencies of negatively charged words (*bad, worst*) including: swear words (*bitch*, *fucking*), words referring to feelings of distress (*mad, cry, hurt, sad*), and to physical complaints (*sick, pain, surgery, hospital*). Notably and in correspondence with the previous analysis, explicit suicide-related word, such as *kill*, *die*, or *suicide* were not included in this list.

In contrast, non-suicidal users who were identified correctly by the MTM (True Negative) had high frequencies of positive words (*great*, *happy*, *perfect*), including positive emotions (*loving*, *love, peace*) and events (*wedding, thanksgiving*), positive experiences of belonging and friendships (*together, friends, mother, wife*), and positive attitude towards life (*blessed, gift, wishes*). Curiously, a dominant theme in the postings of non-suicidal users was religion and spirituality (*Christ, church, God, faith*). These findings suggest that the current ANN model does not rely on explicit manifestations of suicide, but on a wide range of textual contents including emotionally-charged (positive vs. negative) topics.

**Discussion**

This research explored whether suicide risk can be predicted from textual Facebook postings. The results from the STM confirmed our first hypothesis (H1) that Facebook texts may predict general and high suicide risk, particularly when the model is applied among relatively active users. The results from the MTM confirmed our second hypothesis (H2): When the prediction algorithm incorporated a theory-driven hierarchy of psychosocial variables relevant to suicide risk, the quality of the prediction improved substantially, resulting in a medium-to-large and large effect sizes for high and general suicide risk, respectively. The strength of the predictions (.690 ≥ AUC ≤ .759) matched, and sometimes surpassed, previously reported measures in related studies that predicted other psychiatric conditions (e.g., depression, PTSD) from social media (for a review see: 5).

***Theoretical Contributions to Research on Suicide Detection from Social Network Activity***

This research builds on earlier attempts to predict suicide risk from social media by incorporating several improvements. First and most importantly, we collected external, clinically valid measures of suicide risk instead of relying on *proxy diagnostic signals* (posts with explicit references to suicide) (10, 12, 13, 14). Additionally, we collected external measures on psychiatric and psychosocial variables known to contribute to suicide risk. Incorporation of these theory-driven measures insured the construct and external validity of the findings and contributed significantly to the improvement of the predictions. This is noteworthy because most previous studies focused on one psychiatric phenomenon without considering its wider theoretical framework.

Second, the dataset on which the prediction algorithms were developed was meticulously constructed to be of high quality, and is, to the best of our knowledge, the largest of its kind (5). A strict data quality assurance protocol was applied to make sure that only valid responses were included and *post hoc* internal reliability and convergence validity checks were conducted on all variables (see Supplementary Information).

Third, to the best of our knowledge, this study is the first to apply state-of-the-art artificial neural networks and deep contextualized embeddings for text representations in the context of suicide risk prediction from social media. The current use of ELMo has two advantages over other word embedding techniques, such as word count or N-grams. It provides vectors also to non-words popular in social media language (e.g., Lolll or OMG) and enables representations of words within their context (i.e., a given word can receive different vectors, depending on its place in the text).

Fourth, the various procedures of the study including configuration of ANN models, reliance on external measures for suicide instead of explicit suicidal postings, and focusing on everyday language from this popular social network, allowed the extraction of valuable patterns, which could not be hypothesized *a priori*. Algorithms that rely solely on explicit distress-related content could produce False Negative results. In contrast, the proposed models can detect subtler digital footprints of mental health difficulties. Our word search for explicit suicide references revealed that the majority of the users who were identified to be at risk rarely posted content that directly referred to suicide. Correspondingly, the TF-IDFanalysis did not reveal explicit suicide-related words.

Although interpretations remain speculative, the TF-IDF outcomes suggest that correct classifications of suicide risk (True Positive) could be based on high frequencies of negatively charged words (swearing, distress, physical complaints). These negative themes are in line with previous work on digital footprints of depression in social media activity (7). It is also possible that the correct classification considered the language used by the non-suicidal users (True Negative), which included references to positive emotions and experiences, positive attitudes towards life, and religion and spirituality. This is in line with previous work emphasizing the role of meaning in life and religious/community involvement as important protecting factors against actual suicide behaviors (43). Thus, we encourage future researchers to use ANN multi-task models, which could detect suicidal users even when they do not share explicit, suicide-related content.

***Limitations of the Current Research***

The main limitation of the present work concerns the self-report nature of the psycho-diagnostic data collection procedure. Although use of such screening tools is common in large-scale mental health surveys, they cannot match the precision and detailed diagnosis of formal medical assessments of suicide risk (or related psychiatric disorders) by trained mental health experts in face-to-face, clinical interviews. In this study, we chose well-established psycho-diagnostic measures and ensured the quality of the self-reported responses by using multiple validation checks (internal reliability, convergence validity, and a data quality assurance protocol; see Supplementary Information). Nevertheless, we recommend that future research include additional forms of external criteria for suicide risk assessment.

Another limitation concerns the focus on language-based input to the ANN models. A recent study on depression detection indicated the superiority of textual contents over other types of social network signals, such as length or timestamps of postings (7). It is possible however, that additional social network features not included in the current research or in previous studies could potentially improve suicide risk predictions (e.g., reactions to posts, images, videos). Research that will incorporate additional non-textual inputs may improve the quality of suicide risk predictions.

***Implications of the Current Research***

This study has potential implications for the development of practical, effective suicide risk detection tools. The use of AUC scores allowed us to estimate prediction qualities, without the need to establish a pre-defined threshold for flagging a given user as at risk. This has practical implications for suicide predictions because the exact threshold for suicide risk may vary between different end users of such a tool. Some operators prefer a cautious threshold of suicide risk that avoids false alarms (False Positive), while others prefer a sensitive model that identifies as many potentially suicidal individuals as possible (True Positive), even at the expense of some false alarms. These tools could contribute to global efforts to reduce suicide rates by improving early risk detection, both among individuals already receiving mental health care and among the many who do not.

A second implication relate to researchers in computational psychiatry. The use of computational methods to study psychological and psychiatric phenomena from and everyday online activities is becoming increasingly popular. Based on the current findings, we recommend that such endeavors combine state-of-the-art techniques and theory-driven components from clinical and social sciences. While this study did not include every known risk factor, it anchored the predictions of suicide risk within the theoretical framework of the multifaceted nature of suicide (2). We evidenced significant improvements in suicide risk predictions when the detection algorithms were developed based on models that included the wider clinical picture of suicide and its related psychiatric and psychosocial risk factors. In the present study, this progress was made possible due to close collaboration between computational, social, and clinical scientists. Genuine, multi-disciplinary collaboration seems to be a prerequisite for the field of computational mental health research to make significant progress.

**References**

1. Abubakar I, Tillmann T, Banerjee A. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015; 385(9963):117–71.
2. Levi-Belz Y, Gvion Y, Apter A. Editorial: The psychology of suicide: from research understandings to intervention and treatment. Front Psychiatry. 2019; 10:214.
3. Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. CLPsych 2015 shared task: depression and PTSD on Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality; 2015. p. 31–9.
4. De Choudhury M, Counts S, Horvitz EJ, Hoff A. Characterizing and predicting postpartum depression from shared Facebook data. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing; 2014, p. 626–38.
5. Guntuku SC, Yaden DB, Kern ML, Ungar, LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci. 2017;18:43-49.
6. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In: Seventh International AAAI Conference on Weblogs and Social Media; 2013 June. p. 1–10.
7. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro et al. Facebook language predicts depression in medical records. Proc Natl Acad Sci of U S A. 2018;115(44): 11203–8.
8. Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EL. Forecasting the onset and course of mental illness with Twitter data. 2016, August 27. arXiv:1608.07740.
9. Hawton K, van Heeringen K. Suicide. Lancet. 2009;373(9672):1372–81.
10. Homan CM, Johar R, Liu T, Lytle M, Silenzio V, Alm CO. Toward macro-insights for suicide prevention: analyzing fine-grained distress at scale. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality; 2014; Baltimore, MD: Association for Computational Linguistics; 2014. p. 107–17.
11. Niederhoffer K, Hollingshead K, Resnik P, Resnik R, Loveys K. Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology.Stroudsberg, PA: Association for Computational Linguistics; 2019.
12. Sawhney R, Manchanda P, Singh R, Aggarwal S. A computational approach to feature extraction for identification of suicidal ideation in tweets. In Proceedings of ACL 2018 Student Research Workshop; 2018. p. 91–8.
13. Zirikly A, Resnik P, Uzuner O, Hollingshead K. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology; 2019. p. 24–33.
14. Ernala SK, Birnbaum ML, Candan KA, Rizvi AF, Sterling WA, Kane JM, et al. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; 2019 May; p. 1-16.
15. Posner K, Brown GK, Stanley B., Brent DA, Yershova KV, Oquendo MA, et al. The Columbia–Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry. 2011; 168(12):1266–77.
16. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). Washington, DC: American Psychiatric Association; 2013.
17. Sartorius N, Üstün TB, Lecrubier Y, Wittchen HU. (1996). Depression comorbid with anxiety: results from the WHO study on psychological disorders in primary health care. Br J Psychiatry. 1996;168(S30):38–43.
18. Beck AT. Cognitive therapy: a 30-year retrospective. Am Psychol. 1991; 46(4), 368.
19. Nolen-Hoeksema S, Watkins ER. A heuristic for developing transdiagnostic models of psychopathology: explaining multifinality and divergent trajectories. Perspect Psychol Sci, 2011; 6(6), 589–609.
20. Ehring T, Watkins ER. Repetitive negative thinking as a transdiagnostic process. Int J Cogn Ther. 2008;1(3):192–205.
21. Cacioppo JY, Hughes ME, Waite LJ, Hawkley LC, Thisted RA. Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses. Psychol Aging. 2006; 21(1):140.
22. Green BH, Copeland JRM, Dewey ME, Sharma V, Saunders PA, Davidson, I, et al. Risk factors for depression in elderly people: a prospective study. Acta Psychiatr Scand. 1992;86(3):213–7.
23. John OP, Srivastava S. The Big Five Trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin LA, John OP, editors. Handbook of personality: Theory and research. Second edition. New York: Guilford Press; 1999. p. 102–38.
24. Brezo J, Joel P, Gustavo T. Personality traits as correlates of suicidal ideation, suicide attempts, and suicide completions: a systematic review. Acta Psychiatr Scand. 2006; 113(3):180–206.
25. Drapeau CW, Nadorff MR, McCall WV, Titus CE, Barclay N, Payne, A. Screening for suicide risk in adult sleep patients. Sleep Med Rev. 2019; 46:17–26.
26. Weber AN, Michail M, Thompson A, Fiedorowicz JG. Psychiatric emergencies: Assessing and managing suicidal ideation. Med Clin North Am. 2017;101(3):553–71.
27. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med, 2001;16(9):606–13.
28. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006;166(10):1092–7.
29. Nolen-Hoeksema S, Morrow JA. Prospective study of depression and posttraumatic stress symptoms after a natural disaster: The 1989 Loma Prieta earthquake. J Pers Soc Psychol. 1991; 61(1): 115–21.
30. Meyer TJ, Miller ML, Metzger RL, Borkovec TD. Development and validation of the Penn State Worry Questionnaire. Behav Res Ther. 1990; 28(6):487–95.
31. Russell, DW. UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure. J Pers Assess. 1996; *66*(1):20–40.
32. Diener ED, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess. 1985; 49(1):71–5.
33. Rammstedt B, John, OP. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*. 2007; 41(1):203–12.
34. Ophir Y, Sisso L, Asterhan CSC, Tikochinski R, Reichart R. The turker blues: Hidden factors behind increased depression rates among Amazon’s mechanical turkers. Clin Psychol Sci. 2019, Oct. 2. https://doi.org/10.1177/2167702619865973.
35. Prims JP, Sisso I, Bai H. Suspicious IP online flagging tool. 2019 Oct. 28. Available from: https://itaysisso.shinyapps.io/Bots.
36. Arditte KA, Çek D, Shaw AM, Timpano KR. The importance of assessing clinical phenomena in Mechanical Turk research. Psychol Assess.2016; 28: 684.
37. McCredie MN, Morey LC. Who are the turkers? A characterization of MTurk workers using the personality assessment inventory.Assessment. 2018; 1073191118760709.
38. Peters ME, Neumann M, Lyyer M, Gardner M, Clark, C, Lee K, et al. Deep contextualized word representations. 2018; arXiv preprint arXiv:1802.05365.
39. Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data--recommendations for the use of performance metrics. In: 2013 Humaine association conference on affective computing and intelligent interaction; IEEE; 2013, Sept. p. 245-251.
40. Salgado JF. Transforming the area under the normal curve (AUC) into Cohen’s d, Pearson’s rpb, odds-ratio, and natural log odds-ratio: two conversion tables. European Journal of Psychology Applied to Legal Context, 2018; 10(1):35–47.
41. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, 1: 4171–4186.
42. Mogotsi IC. Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Information Retrieval. 2010; 13:192–5.
43. VanderWeele TJ, Li S, Tsai AC, Kawachi I. Association between religious service attendance and lower suicide rates among US women. JAMA Psychiatry. 2016; 73, 845-51.

**Figures and Tables**

**Figure 1.** The Single Task Model (STM).



Note: FC layers = Fully Connected layers.

**Figure 2.** The Multi Task Model (MTM).



Note: FC layers = Fully Connected layers; The sign ⊕ symbolizes the vector concatenation operator.

**Table 1.** Detection performance (AUC scores) of STM and MTM across all users (*N* = 1,650) and for active users only (*N* = 1,002).

|  |  |  |
| --- | --- | --- |
| Task | General suicide risk | High suicide risk |
| Model | STM | MTM | STM | MTM |
| AUC for All users  | .567 | .602 | .555 | .571 |
| AUC for Active users  | .608 | .759 | .606 | .690 |

Note: STM = Single Task Model; MTM = Multiple Tasks Model; AUC = Area Under the receiver operating characteristic Curve.

1. The 'general risk' group was chosen for the content analyses because it is larger than the sub-group of high-risk individuals, and therefore provides more textual content for interpretation. [↑](#footnote-ref-1)