**Automated Detection of Suicide Risk from Textual Facebook Activity: Data, Neural Network Models, and Language Usage Analysis**

\*Yaakov Ophir,1,2, Refael Tikochinski,1 Christa Asterhan,1 Itay Sisso,1 Roi Reichart2

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

**Author Note**

\*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**

Can Facebook postings reveal suicide risk of users? This article presents Artificial Neural Network (ANN) models designed to predict general and high suicide risk from a large, high-quality dataset. The dataset included 85,643 Facebook postings matched with clinically valid psychiatric and psychosocial information about 1,650 authenticated users. Using ELMo for creating text representations, two principal models were constructed: A Single-Task Model (Facebook content → suicide) confirmed that suicide risk can be predicted from Facebook postings, especially among active Facebook users. A Multi-Task Model, which included hierarchical multilayered sets of theory-driven risk factors (Facebook content → personality traits → psychosocial risks → psychiatric disorders → suicide), produced improved predictions (0.690 ≤ *AUC* ≤ 0.759) with substantially larger effect sizes (0.701 ≤ *Cohen’s d* ≤ 0.994). This increase in the quality of the predictions suggests that machine-learning algorithms may benefit significantly from incorporating theory driven components. Complementary qualitative analyses suggest that the ANN models do not rely on explicit manifestations of suicide (e.g., “I want to kill myself”) but on a wide range of textual contents, including emotionally charged (positive vs. negative) topics. The advanced methodologies of the current study contribute both to the construct validity (i.e., the theoretical framework of suicide) and the ecological validity of the findings (i.e., their applicability in real-life settings), even when users do not publish explicit, distress-related contents. By laying the ground for the development of a practical suicide risk screening tool, this study joins the world-wide efforts to combat and reduce suicide rates.

**Keywords**: suicide; depression, artificial neural network; machine learning, social networking sites

**1. Introduction**

Suicide is a leading cause of death worldwide and early detection of suicide risk is a major priority for improving prevention efforts (Abubakar, Tillmann, & Banerjee, 2015; Levi-Belz, Gvion, & Apter, 2019). Remarkably, new opportunities for suicide risk detection have emerged with the rise of Social Networking Technologies. Billions of people create and upload content to Facebook, Twitter and other social networks (Statista, 2019) that could contain valuable information regarding their psychiatric and emotional functioning (Guntuku, Yaden, Kern, Ungar, & Eichstaedt, 2017). Correspondingly, recent studies in the developing field of computational psychiatry have shown that features of online behavior can predict personality traits and mental health conditions of users (e.g., Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015; De Choudhury, Counts, Horvitz, & Hoff, 2014), including the user tendency to experience major depression (De Choudhury, Gamon, Counts, & Horvitz, 2013; Eichstaedt et al., 2018; Reece et al., 2016), a dominant risk factor for suicide behaviors (American Psychiatric Association, 2013; Hawton & van Heeringen, 2009).

Yet, to date, the existing research on predicting suicide risk from online behavior has relied predominantly on *proxy diagnostic signals*, that is: “tweets” or posts with explicit suicide-related content, which served as the predictive criteria (i.e., “ground truths”) of the detection model (Hawton & van Heeringen, 2009; Homan et al., 2014). For example, the ground truth criterion in the latest workshop in Computational Linguistics and Clinical Psychology (CLPsych 2019; Sawhney, Manchanda, Singh, & Aggarwal, 2018), which dedicated a “shared task” to suicide prediction, was annotated postings that were judged by suicide experts or by non-experts crowdsourcing workers as references to suicidal thoughts or behaviors of the user. Specifically, these annotated postings were collected from a designated, internet-based discussion and support forum on the topic of suicide (Niederhoffer, Hollingshead, Resnik, Resnik, & Loveys, 2019).

Without external clinically valid measures of suicidal thoughts and behaviors, however, existing research on suicide detection from online behaviors suffers from low ecological (external) validity and from low construct validity. The degree to which the findings can be generalized to real life settings (i.e., ecological validity) is limited because the findings can only relate to users who publish explicit references to suicidal behavior, usually in designated online support communities and cannot be generalized to large populations who are active in ubiquitous social networking sites. More importantly, the degree to which the existing studies measure what they claim to be measuring (i.e., construct validity) is limited because online, explicit references to suicide-related topics may not indicate actual suicidal thoughts or behaviors. Moreover, whether online or offline, explicit suicide references are only a small part of a much more complicated psychosocial phenomenon, which has multiple facilitating and contributing risk factors (Levi-Belz et al., 2019).

The overall goal of the present study is to predict suicide risk from social media textual postings, while maintaining high levels of construct and ecological research validity. In order to overcome the validity challenges, we applied two complementary methodological strategies: We collected psychosocial data (ground truths) from 1,650 Facebook users, including suicide risk, psychiatric disorders, psychosocial risks, and personality traits, using clinically valid psycho-diagnostic questionnaires and we documented a full year of Facebook activity among these users, resulting in a dataset of 85,643 Facebook postings. To our knowledge, no other research paper reports on a dataset that consists of social media activity records matched with clinically valid suicide measures, particularly not a dataset of the size we report here (Guntuku et al., 2017). Based on this dataset, we constructed Artificial Neural Network (ANN) models that aimed to predict suicide risk from representations of Facebook texts, which are extracted by the ELMo contextualized word embedding model (Peters et al., 2018; see the Method section).

ANN-based models are especially fit to the task at hand (i.e., predicting suicide from Facebook postings) as they can extract “bottom-up” features that do not include explicit references to suicide. Moreover, they provide a simple and effective platform for learning multiple variables jointly (Ruder, 2017), thus enabling the analysis of a multilayered psychosocial profile of suicidal individuals. Finally, by using ELMo to create representations of texts, ANN models can also consider non-word inputs, which often appear in social media language (e.g., Mmhhmm, Aahhhaa, shiiiiiit, etc.). A full description of the ANN models is provided in the Method section.

Our first hypothesis (H1) was that the content of Facebook postings would include significant signals that could be used for the prediction of suicide risk among Facebook users (Facebook content → suicide). Following the modular structure of the suicide rating scale (Posner et al., 2011) described in the Method section, we conducted Single Task Models (STMs) that aimed to identify two groups of users, a group of users at *general risk of suicide* and a sub-group of users at *high risk of suicide* (i.e., a sub-group of the former “general risk” group).

In light of the fact that suicide predictions (i.e., the model outcomes) are generated based on signals extracted from the Facebook textual content (i.e., the input of the model), we expected that the quality of the suicide predictions will improve among *Active Facebook Users*, compared with less active users. We therefore applied each model in two setups, one that included the entire sample (*N* = 1,650) and one that included only active users (*N* = 1,002) who published a number of postings that was equal to, or higher than, the median number of postings in the sample (*MD* = 10).

In light of the multifaceted nature of suicide, we also collected clinically validated data on three additional sets of risk factors for suicide and for depressive episodes, which in many cases precede the suicide behavior itself (American Psychiatric Association, 2013). These factors are organized in a hierarchical order, according to their decreasing level of severity: (A) psychiatric disorders, (B) psychosocial risk factors and (C) personality traits. The most severe risk factor for suicide is the existence of a psychiatric disorder (Hawton & van Heeringen, 2009). Specifically in this study, we assessed major depressive disorder, which is a proximal predictor of suicide (American Psychiatric Association, 2013), and generalized anxiety disorder (GAD), which often appears in comorbidity with major depression (American Psychiatric Association, 2013; Sartorius, Üstün, Lecrubier, & Wittchen, 1996). The next two sets of factors included variables that are known to predict or to maintain depressive episodes. This is because suicide is usually seen as part of the clinical picture of depression (American Psychiatric Association, 2013; Beck, 1991). The second set of factors included the following theory-driven *psychosocial risks* for depression (Beck, 1991; Nolen-Hoeksema & Watkins, 2011): depressive rumination (brooding), excessive worries (Ehring & Watkins, 2008; Nolen-Hoeksema & Watkins, 2011), feelings of loneliness, and lack of satisfaction with life (Cacioppo, Hughes, Waite, Hawkley, & Thisted, 2006; Green et al., 1992). The third set of factors included the “Big Five” personality traits (John & Srivastava, 1999) that can be relevant to depression and to suicide prediction, especially in the case of neuroticism (and to a lesser extent, extroversion), which in combination with stressful life events is associated with depressive symptoms (American Psychiatric Association, 2013) and suicide behaviors (Brezo, Joel, & Gustavo, 2006). Based on this theoretical view of suicide and depression, our second hypothesis (H2) was that a Multi Task Model (MTM) that aims to predict suicide risk, while taking into account all three layers of contributing factors (Facebook content → personality traits → psychosocial risks → psychiatric disorders → suicide), would yield improved predictions of general/high suicide risk compared with the straightforward Single Task Model (Facebook content → suicide).

Finally, we provide a human interpretation for the computational “black box” via complementary qualitative analyses that explore the specific textual indications that perhaps allowed the machine to make its predictions. By offering human interpretation and through the examination of the two hypotheses above, the current work is expected to contribute to the existing research on suicide risk detection from online behavior as follows: (a) The use of a clinically valid tool for measuring suicidal thoughts and intent (with or without a specific method or a concrete plan), alongside (b) other clinically valid measures that assess a wide range of psychiatric and psychosocial risk factors of 1,650 users, while (c) taking into account their level of activity on Facebook, and (d) analyzing the textual themes of their postings, strengthen both the construct validity and the ecological validity of suicide detection efforts. By strengthening the validity of suicide prediction, we hope that the findings of the current study could be generalized to large populations and to various social networks, and that they will accelerate the development of practical suicide risk detection tools.

1. **Method**
	1. **Tools and measurements**

**Facebook data collection.** A designated Facebook application was developed for the purpose of the current research. Similar to other popular Facebook apps (e.g., Candy Crush), this application extracts data from social media to external data storage, upon the users’ explicit authorization. By agreeing to participate in the study and by installing the designated application, participants (*N* = 1,650) gave us a one-time authorization to download their Facebook status updates (i.e., Facebook posts) from the past year prior to the beginning of the study. Altogether, the application extracted a total of 85,643 Facebook posts, which were published on the user public timeline. For the purpose of this study, we were only interested in original postings that were generated by the participants themselves (i.e., not other people’s postings, which were “shared” by the user). The *median* number of Facebook postings per profile was 10 (*M* = 42.99, *SD* = 86.28) and the *median* number of words in each post was 27 (*M* = 35.23, *SD* = 38.42). A total of 1,002 participants published at least 10 posts and were therefore marked as “*Active Facebook users*.”

**Suicide risk.** Suicide risk was measured using the well-established Columbia Suicide Severity Rating Scale (CSSRS; Posner et al., 2011). The CSSRS was originally developed to help clinicians structure their clinical interviews and assess the existence and severity of suicide risk with high levels of accuracy. The scale demonstrated high sensitivity and specificity scores in suicide prediction and it is considered a “diagnostic tool of choice,” both in clinical settings and in empirical research (Drapeau et al., 2019; Weber, Michail, Thompson, & Fiedorowicz, 2017). Upon consultation with the principal developer of the CSSRS (Posner, personal written communication), we chose to administer the electronic self-report version of the scale, in light of the fact that the current research examined participants from a crowdsourcing platform. The electronic version of the CSSRS has been demonstrated to have psychometric validity and prediction accuracies that are comparable to the original clinician version of the scale (Mundt et al., 2010; Viguera et al., 2015).

The scale consists of six binary (yes/no) items that are presented to the participants in two parts. In the first part, participants were asked to complete Item 1 that addressed a “wish to be dead” *(“Have you wished you were dead or wished you could go to sleep and not wake up?”*) and Item 2 that addressed “suicidal thoughts” (*“Have you actually had any thoughts of killing yourself?”*). Only if participants answered “yes” to item #2 on suicidal thoughts, they were then exposed to the second part of the scale that examined the severity of the risk. Item 3 addressed suicidal thoughts with method (*“Have you been thinking about how you might kill yourself?”*). Item 4 addressed suicidal intent (*“Have you had these thoughts and had some intention of acting on them?”*). Item 5 addressed suicide intent with specific plan (*“Have you started to work out or worked out the details of how to kill yourself? Do you intend to carry out this plan?”*), and Item 6 addressed actual suicide behaviors (*“Have you ever done anything, started to do anything, or prepared to do anything to end your life?”*). Participants who answered “yes” to this last item were then asked to indicate when they engaged in such behavior (over a year ago, between three months and a year ago, or within the last three months).

The modular structure of the CSSRS enables the extraction of two binary (yes/no) variables: a *general risk of suicide* (participants who met the criterion of the first part of the scale, that is answering “yes” to item 2) and a *high risk of suicide* (a sub-group of participants at suicide risk who also responded “yes” to at least one of the items in the second part of the scale). The total sum score of the “yes” answers to all six items serves as another indication for the severity of the suicide risk.In this study, the total score of the CSSRS was positively correlated with all the examined risk factors (Table 1) and especially with depression (*r* = 0.46), thus indicating a high convergent validity of the scale (for further details on the convergent validity of the various scales, see the *Supplementary Information*).

**Risk factors for suicide and depression.** A detailed description of all the psycho-diagnostic measurements that were used in addition to the suicide scale is provided in the *Supplementary Information*. In short, major depressive disorder was measured using the Patient Health Questionnaire-9 (PHQ-9; Kroenke, Spitzer, & Williams, 2001). Generalized anxiety disorder was measured using the GAD-7 (Spitzer, Kroenke, Williams, & Löwe, 2006). Depressive rumination (Brooding) was measured using five items from the Ruminative Responses Scale (RSS; Nolen-Hoeksema & Morrow, 1991). Excessive worrying was measured using the Penn State Worry Questionnaire (PSWQ; Meyer, Miller, Metzger, & Borkovec, 1990). Loneliness was measured using the 10-item version of the UCLA-Loneliness Scale (Russell, 1996). Low satisfaction with life was measured using the Satisfaction With Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985). Finally, personality traits were assessed using the short version of the Big Five Inventory (BFI-10; Rammstedt & John, 2007). The convergent validity of the various psychosocial scales was high. For further information about the scales, see the *Supplementary information*.

* 1. **Dataset**

The procedure of the study has been approved by the Ethics for Research on Human Subjects Committees at both the Technion Israel Institute of Technology and of the Hebrew University of Jerusalem. The recruitment of research participants was conducted through Amazon’s Mechanical Turk (MTurk), a widely used crowdsourcing platform online (Buhrmester, Kwang, & Gosling, 2011). Inclusion criteria were: owning a Facebook account and having previous proven experience in MTurk-based studies. Proven experience was defined as past completions of at least 100 MTurk tasks, with a minimum of 95% success rate. The recruited participants read and signed a consent form that described the requirements of the study in detail. They then continued to complete the eight self-report psycho-diagnostic measures mentioned above and were asked to install a designated Facebook application that extracted their Facebook content to external data storage. Upon completion of the study, participants who met the criterion for suicide risk received a designated letter that included a list of mental health services and an encouragement to seek help (see the *Supplementary Information* for the complete ethical research-protocol). All participants were compensated for their participation with a sum of $2.

In light of recent concerns regarding the quality of crowdsourcing-based data, we applied a newly developed rigid data quality assurance protocol (Ophir, Sisso, Asterhan, Tikochinski, & Reichart, 2019). To avoid bogus responses, we limited the participation to US residents and excluded users with suspicious Internet Protocol (IP) addresses (Prims, Sisso, & Bai, 2018). To ensure the quality of the unsupervised responses, we implemented a designated inattentiveness scale that comprised eight hidden attention checks. These checks included four types of data-quality measurements (i.e., “infrequency items,” “time measurements,” “person-total correlation,” and “long string analysis”), which were embedded in the various self-report scales of the study (Ophir, Sisso, et al., 2019).

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 or non-US IP addresses. From the remaining users, 1,985 users were marked as valid users who passed the eight attention checks successfully. A total of 335 users did not publish any textual status updates (posts) on their Facebook account and therefore could not be included in the current study that focused on predictions from textual contents. The final sample of attentive participants who published at least one Facebook post was therefore 1,650.

Descriptive statistics and zero order correlations of the psycho-diagnostic measures are presented in Table 1. Based on previous works and on the psychological compositions of MTurk samples, we note that the prevalence of mental health issues and especially of major depression is significantly higher in MTurk, compared with the general population (Arditte, Çek, Shaw, & Timpano, 2016; McCredie & Morey, 2018; Ophir, Sisso, et al., 2019). Correspondingly, in the current sample, we evidenced high rates of suicide risk. A total of 568 users (34.4%) met the criterion of a *general risk of suicide*, of which 204 (12.4%) met the criterion of *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.

* 1. **ANN-based Models**

Two ANN-based models were constructed (Figures 1 and 2). The architectures of both models consisted of identical input and output layers. The input fed into the models consisted of representations of Facebook texts, which are 1024-dimensional vectors extracted by the ELMo contextualized word embeddings model (Peters et al., 2018; see next). 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, where the variants differ in their output variable (i.e., *a general risk of suicide* and a *high risk of suicide*).

The two variants of the Single Task Model (STM) aimed to predict suicide risk directly from Facebook activities (Facebook content → suicide), without the inclusion of other risk factors. The two variants of the Multi-Task Model (MTM) aimed to predict a hierarchical combination of multiple factors. Based on the literature on suicide and depression presented above, we constructed a hierarchical “pyramid” of risk factors starting from *suicide risk* at the top and expanding to relevant *psychiatric disorders*, *psychosocial risk* factors and *personality traits* at the bottom (Figure 3).

* 1. **The architectures of the two models**

As illustrated in Figure 1, the Single-Task Model (STM) consisted of an input and an output layer, which are connected by a set of fully-connected layers. In contrast, the Multi-Task Model (MTM) contained three additional hierarchically organized auxiliary layers: Facebook content → personality traits → psychosocial risks → psychiatric disorders → suicide risk. As illustrated in Figure 2, each auxiliary layer is accompanied by a set of fully-connected layers, thus forming several “subnetworks.”

The subnetwork located at the bottom of the model (i.e., the Personality traits) is activated directly by the input layer (Facebook content), while the subnetworks at the middle (Psychiatric disorders and Psychosocial risks) are activated by the previous subnetwork’s output, which is concatenated with outputs from a shared set of fully-connected layers. The shared set of layers is activated directly from the input layer and allows the subnetworks to get direct information from the input layer (and not just from the previous subnetwork). This architecture introduces inductive bias to the suicide prediction model through the auxiliary tasks, while learning a shared set of parameters for the multiple tasks to reduce the risk for overfitting. Finally, the Suicide layer at the top of the model is activated by the output generated from the Psychiatric disorders layer and from the outputs of the shared set of hidden layers (Figure 2).

The loss function of the STM models is the *binary cross-entropy*:

Where *N* is the number of training examples, indicates whether participant *i* belongs to the suicide group () or not () according to the ground truth, and indicates the probability of as predicted by the model.

The loss function of the MTM is the sum of the output layer’s and the auxiliary layers’ loss functions:

Where is the binary cross-entropy loss function like before, and is the sum of all *mean squared errors* (MSEs) calculated for each of the auxiliary variables in the set *A*={Depression, Anxiety, Brooding, Worry, SWL, Lonely, Open, Conscientious, Extravert, Agreeable, Neurotic}:

where *N* is the number of training examples, is a continuous variable representing the ground truth score of the auxiliary-variable *a* for subject *i*, and is the predicted score for this variable according to the model.

The textual content of the Facebook postings was encoded using ELMo, a state-of-the-art ANN framework for “Embeddings from Language Models” (Peters et al., 2018). ELMo comprises a deep language model through multiple bi-directional Long-short-Term-Memory (LSTM) layers. ELMo has been shown to produce contextualized word embeddings that are more effective in many Natural Language Processing (NLP) tasks, compared to state-of-the-art non-contextualized embeddings such as Glove (Pennington, Socher, & Manning, 2014). Furthermore, ELMo is especially relevant to social media language. This is because ELMo is character-based (rather than word-based), thus allowing the system to make representations also to non-words (i.e. words that do not appear in formal dictionaries), as well as to expressions that did not appear in the learning phase. Using a pre-trained ELMo model (available at https://tfhub.dev/google/elmo/2), we extracted a 1024-dimensional embedding vector for each Facebook post in our data through mean-pooling over the contextualized word embeddings generated for the post. The overall textual-activity of the user was represented as the average of its post vectors. The resulting 1024-dimensional vector (per user) was then used as the input to the ANN models.

In the learning phase, each ANN-based model was trained on 70% of the input data (i.e., the Facebook texts of 1,155 users out of the entire sample), so that it could distinguish between Facebook patterns of suicidal and non-suicidal individuals. De facto, each learning example is comprised of the Facebook texts of one of the participants together with one of the suicide labels of the participant (i.e., general/high suicide risk), and for the MTM model it also includes the auxiliary variables scores of this participant. Then, in the development phase, a hyper-parameter tuning process was conducted on another 15% of the data (i.e., 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 only one of the MTM three auxiliary layers (i.e., psychiatric disorders/psychosocial risks/personality traits) and their detection performance did not reach the prediction quality of the complete MTM. Finally, in the test phase, the remaining 15% of the dataset (i.e., 248 users) was used to examine the predictive quality of each model. Further details on this machine learning process, including the optimization of the hyper-parameters of the models, is provided in the *Supplementary Information*.

**3. Results**

* 1. **Detection performance of suicide risk**

The ANN models produced binary (yes/no) predictions regarding the two (general/high) suicide risk variables for each Facebook user. These predictions were categorized into one of four possible classes: True Positive, in which a suicidal user is correctly detected (true) by the model as suicidal (positive); False Positive, in which a non-suicidal user is incorrectly detected (false) as suicidal (positive); True Negative in which a non-suicidal user is correctly determined (true) as not suicidal (negative); and False Negative in which a suicidal user is incorrectly determined by the model (false) as non-suicidal (negative). Following this classification, a Receiver Operating Characteristic curve (ROC curve), which plots the True Positive rates against the False Positive rates was generated and the Area Under the ROC Curve (AUC) was calculated. This last measure (AUC) provides a reliable estimation of the quality of the predictions across all possible classification thresholds and it can be transformed to the common effect size measure (*Cohen’s d*) in experimental psychology (Salgado, 2018).

Table 2 demonstrates the detection performance of the two models for the two types of suicide risk. In support of our first hypothesis (H1), the performance of the Single Task Model (STM) shows that Facebook content includes discernable signals that can be used for the prediction of suicide risk, even when the model is applied to all Facebook users (Table 2), regardless of their activity level (AUC = 0.567 and 0.555, for general and high suicide risk, respectively). However, and as expected, the performance measures improve when the model is applied to Active Facebook Users only (Table 2) (AUC = 0.608 and 0.606 for general and high suicide risk, respectively). A transformation of these AUC scores to *Cohen’s d* scores (Salgado, 2018) indicated a small to medium effect size for general risk (*Cohen’s d* = 0.388) and high risk (*Cohen’s d* = 0.380) of suicide.

Most importantly, the inclusion of all risk factors in one Multiple Task Model (MTM) yielded improved predictions, especially among Active Facebook Users (AUC = 0.759 and 0.690, for general and high suicide risk, respectively) (Table 2). These predictions indicated a medium to large effect for high risk of suicide (*Cohen’s d* = 0.701) and a large effect for general risk of suicide (*Cohen’s d* = 0.994). These 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 compared with a single task model that targets suicide risk only. Future efforts may build on the current findings and incorporate additional theory-driven risk factors when developing suicide risk detection algorithms.

* 1. **Qualitative interpretation of the observed predictions**

In order to offer an initial human interpretation for the observed ANN-based predictions (i.e., to evaluate what could be the specific textual indications that allowed the machine to make predictions on suicide risk), we conducted two post-hoc qualitative content analyses. In the first analysis, we conducted a word-search for explicit suicide manifestations among active users who were classified correctly by the MTM as users at general risk of suicide[[1]](#footnote-1) (“True Positive”) (*N* = 33, 22% of the test data). This search yielded eight appearances of the terms “suicide”/”suicidal,” 20 appearances of “kill,” and 44 appearances of “die” (including “dying,” “dead,” and “death”). Interestingly, none of these explicit wordings (except one) indicated a suicide risk of the users. Two examples for such postings 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.”

In the second analysis we applied *Term Frequency Inverse Document Frequency (TF-IDF)* analysis (Mogotsi, Manning, Raghavan, & Schütze, 2010) to extract the hundred most frequent words that best distinguished between the four classes (i.e., True Positive, True Negative, False Positive, and False Negative) of general suicide risk prediction among active users (Table 3). Interestingly, users at general suicide risk who were identified correctly by the MTM (i.e., True Positive) had high frequencies of negatively charged words (bad, worst) including: swear words (bitch, fucking), words referring to feelings of distress (lose, mad, poor, cry, hurt, sad), and to physical complaints (sick, pain, surgery, hospital). Notably and correspondingly with the previous analysis, explicit, suicide-related wordings, 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 (dear, together, friend, friends, brother, 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 (lord, Christ, church, god, faith, Christmas, Jesus, pray, prayer, prayers, heaven, father, son, spirit). These qualitative 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 (Table 3).

**4. Discussion**

Recent studies in the growing field of computational psychiatry suggest that data from social media contains valuable information about the mental health of the users (Eichstaedt et al., 2018; Guntuku et al., 2017). The overall goal of the current research was to examine whether the textual content of Facebook postings can predict valid suicide risk of users. Specifically, we collected numerous postings (*N* = 85,643) from the preceding year activity of 1,650 Facebook users, and cross-examined their textual content with clinically valid psychiatric and psychosocial diagnostic data. To the best of our knowledge, this procedure formed the largest dataset for clinically valid suicide prediction from social media (Guntuku et al., 2017). The results from the ANN-based, Single Task Model (STM) confirmed our first hypothesis (H1) that Facebook postings can predict general risk and high risk of suicide, especially when the model is applied among active Facebook users. The results from the Multiple Task Model (MTM) confirmed our second hypothesis (H2) that a multilayered prediction model that explicitly accounts for a hierarchy of psychosocial variables (Facebook content → personality traits → psychosocial risks → psychiatric disorders → suicide) would produce substantially improved predictions, in comparison to a model that targets suicide risk only. The findings on active Facebook users indicated a medium to large effect for high risk of suicide (*Cohen’s d* = 0.701) and a large effect for general risk of suicide (*Cohen’s d* = 0.994; Salgado, 2018). The quality of the predictions (0.690 ≥ AUC ≤ 0.759) matched, and in some cases surpassed, previously reported measures in related studies that predicted other psychiatric conditions (e.g., depression, PTSD) from social media (Guntuku et al., 2017).

**4.1. Theoretical contributions to the validity of suicide detection from social networks**

Perhaps due to the complexity of collecting valid psychological data and the inherent ethical challenges of suicide research, previous works on suicide detection from social networking technologies did not include clinically-valid measures of suicide risk, as well as other psychosocial characteristics of social media users (Homan et al., 2014; Sawhney et al., 2018; Zirikly, Resnik, Uzuner, & Hollingshead, 2019). The current study addressed this gap in the literature by forming a large high-quality dataset and applying advanced language processing strategies, which improve the construct validity and the ecological validity of suicide detection from social media.

The *construct validity* of the current study’s findings is high, especially in light of the valid ground truths regarding the user suicidality, which are used in both ANN training and evaluation. Moreover, the administration of multiple well-established clinical assessment tools alongside the configuration of a multi task ANN-based model (MTM) anchors the suicide predictions within the theoretical framework of the multifaceted nature of suicide and its potential risk factors (Levi-Belz et al., 2019). As opposed to linear classifiers, ANN models are very flexible in modeling multiple variables jointly (Ruder, 2017). In the current study, a hierarchical model of risks associated with suicide risk was proposed (Figure 3). While this hypothetical model does not include each and every risk factor that possibly exists, it serves as a practical aid for accurate detection of suicide. In a way, the deep learning model can be seen as an artificial imitation of human-based deliberations made by psychiatrists conducting suicide assessment.

The *ecological validity* of the findings is also high, in light of the state-of-the-art language processing and machine learning techniques that were used: the framework of ELMo, which was applied in order to create input text representations, and the ANN-based model that consisted of multiple layers of psychiatric and psychosocial risk factors of suicide (MTM). In contrast to other word embedding techniques such as N-grams, bag of words, and even GLove, ELMo creates representations for words within their context (i.e., a given word can receive different vectors, depending on its place in the text). Moreover, in light of the fact that ELMo is character-based (rather than word-based), it provides vectors also to non-words that are popular in social media language (e.g., Lollll, OMG, etc.).

Complementing this approach, the configuration of the ANN-based model allows the extraction of bottom-up patterns, which might not be hypothesized a priory. This is noteworthy, because the reliance on explicit distress-related contents as a sole top-down predictor could miss many users who prefer not to share their emotional distress explicitly online (Ophir, Asterhan, & Schwarz, 2019). In fact, our own word-search for explicit suicide manifestations in the test data revealed that the vast majority of the users did not publish explicit wishes to engage in suicide behavior. As shown above, except one explicit suicide manifestation, all the explicit wordings (e.g., “die,” “kill,” “suicide”) were not related to suicidal thoughts or behaviors of the Facebook user. Correspondingly, the *TF-IDF* analysis (Table 3) did not reveal explicit suicide-related words.

It is hard of course to provide interpretation of the “black box” and evaluate what were the specific textual indications that allowed the machine to make predictions on suicide risk. Yet the above analysis (*TF-IDF*), which extracted the hundred most frequent words that best distinguished between the four classes of prediction (Table 3), suggests a possible interpretation according to which correct classifications of suicide risk (i.e., True Positive) were made based on high frequencies of negatively charged words (i.e., swears, distress, and physical complaints). These negative themes correspond with previous works that addressed depression manifestations in social media (Eichstaedt et al., 2018; Ophir, Asterhan, & Schwarz, 2017). It is also possible that the correct classification took into account the language used by the non-suicidal users (True Negative) who had high frequencies of positive emotions and experiences alongside positive attitudes towards life and multiple references to religion and spirituality. These findings are in line with previous literature that singled meaning in life and religious involvement as important protecting factors against actual suicide behaviors (Stack, 1983; VanderWeele, Li, Tsai, & Kawachi, 2016).

Altogether, the religious words along with the emotionally charged (positive vs. negative) topics add another contribution to the construct validity of suicide predictions and provide a preliminary interpretation of the digital footprints that might underlie the observed predictions. In light of this qualitative interpretation, we encourage researchers to apply ANN multi-task models, which are based on contextualized word embeddings for textual representations, when they try to detect mental health conditions from social media. By using these computational strategies, researchers might be able to detect suicidal users, even in cases where these users did not publish any explicit, suicide-related contents.

Three additional characteristics of the current dataset contribute to the emerging field of computational psychiatry: the recruitment of a large sample, the application of a strict data quality assurance protocol (Ophir, Sisso, et al., 2019), and the focus on the social network of Facebook. First, in light of the fact that machine learning models require a large number of examples to learn from (in our case, the Facebook postings and the various psychosocial characteristics of each user), the collection of data from a large sample of users allowed the achievement of high quality predictions. Second, many of the previous studies that did use valid clinical tools (e.g., studies that predicted depression) examined small to medium sized research samples (200 ≤ *N* ≤ 700; Guntuku et al., 2017) with limited quality controls. In contrast, the current study is based on a large, high-quality dataset, which contributes to the reliability of the results. Finally, previous works have mainly focused on Twitter (Guntuku et al., 2017), where more permissive privacy protection policies provide researchers with easy data collection opportunities. However, Facebook is currently the most popular social networking technology, with an estimated number of two billion users (Statista, 2019). Therefore, the current choice to target Facebook enlarges the size of the population that could potentially benefit from a suicide risk detection algorithm and contributes to the accumulating research that aims to develop automated and powerful tools for early diagnosis of various psychiatric conditions (Corcoran et al., 2018).

**4.2. Limitations of the current research**

The current study has limitations. Even though our main efforts were directed at the improvement of the validity of suicide detection studies, the psycho-diagnostic tools used in this study relied on self-report responses from participants. Indeed, the use of such tools is rather common in large scale mental health surveys (Centers for Disease Control and Prevention [CDC] & National Center for Health Statistics, 2018). Yet, these screening tools have limitations and they cannot replace formal medical assessments of suicide risk or of psychiatric disorders that are determined by trained psychiatrists in face-to-face, clinical interviews. In the current study, we specifically chose well-established psycho-diagnostic measures and ensured their validity using other validation checks (internal reliability, convergence validity, and a data quality assurance protocol), described in detail in previous work (Ophir, Sisso, et al., 2019). However, we recommend that future studies will also include face-to-face clinical interviews with Facebook users, which could further investigate the validity of the current findings.

The second limitation concerns the focus on language-based input to the ANN models. Even though a recent study on depression detection has shown the superiority of textual contents over other types of social network signals, such as length or timestamps of postings (Eichstaedt et al., 2018), the addition of other Facebook activity features could potentially improve suicide risk predictions. Candidates for potentially relevant features are, among others, non-textual reactions by others to the user’s posts (e.g., “likes,” angry or sad emoji), postings time and frequency, and features of images and videos that were uploaded by the user. Future work should endeavor to include textual as well as non-textual signals. As noted, ANN models are highly suitable for modeling multiple types of input. We therefore expect that extending the current models to include additional input signals would yield effective models.

**4.3.** **Practical implications of the current research**

Despite the above limitations, the findings of the current study can be used as the scientific ground for the development of practical and inexpensive suicide detection tools from online behavior. Aside from the promising results of the current study for future developments, the take home message from its methodology is that machine learning algorithms should incorporate theory driven components. Only when we considered the wide range clinical picture of suicide and its related psychiatric and psychosocial risks, we could suggest human interpretation for the artificial models and reach the high quality predictions of suicide risk among active Facebook users. In fact, the high AUC scores give an overall estimation of the quality of the models without the need to establish a pre-defined threshold for suicide risk. This is especially important in suicide predictions because the threshold for suicide risk may vary according to the operator’s needs. Some operators of the algorithm might prefer a cautious threshold of suicide risk that avoids false alarms (False Positive) while others might prefer a very sensitive model that identifies as many suicidal individuals as possible (True Positive), even at the expanse of some false alarms. Hopefully, the multifaceted methodology and the promising results of this study could accelerate the development of practical suicide risk screening tools, which could contribute to the global efforts to reduce suicide rates and which will perhaps even save human lives.

**Acknowledgments**

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

**References**

Abubakar, I., Tillmann, T., & Banerjee, A. (2015). 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*, *385*(9963), 117–171.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®*). Washington, DC: Author.

Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment*, *28*, 684.

Beck, A. T. (1991). Cognitive therapy: A 30-year retrospective. *American Psychologist*, *46*(4), 368.

Brezo, J., Joel, P., & Gustavo, T. (2006). Personality traits as correlates of suicidal ideation, suicide attempts, and suicide completions: A systematic review. *Acta Psychiatrica Scandinavica*, *113*(3), 180–206.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3­–5.

Cacioppo, J. Y., Hughes, M. E., Waite, L. J., Hawkley, L. C., & Thisted, R. A. (2006). Loneliness as a specific risk factor for depressive symptoms: Cross-sectional and longitudinal analyses. *Psychology and Aging*, *21*(1), 140.

Centers for Disease Control and Prevention (CDC) & National Center for Health Statistics. (2018). *Data from National Health and Nutrition Examination Survey, 2015–2016*.Washington, DC:U.S. Department of Health and Human Services.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). 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* (pp. 31–39).

Corcoran, C. M., Carrillo, F., Fernández‐Slezak, D., Bedi, G., Klim, C., Javitt, D. C., … Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*: 67–75. doi:10.1002/wps.20491.

De Choudhury, M., Counts, S., Horvitz, E. J., & Hoff, A. (2014). Characterizing and predicting postpartum depression from shared Facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 626–638).

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *ICWSM*, *13*, 1–10.

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75.

Drapeau, C. W., Nadorff, M. R., McCall, W. V., Titus, C. E., Barclay, N., & Payne, A. (2019). Screening for suicide risk in adult sleep patients. *Sleep Medicine Reviews*, *46*, 17–26.

Ehring, T., & Watkins, E. R. (2008). Repetitive negative thinking as a transdiagnostic process. *International Journal of Cognitive Therapy*, *1*(3), 192–205.

Eichstaedt, J. C., Smith, R. J.,  Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, … Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(44), 11203–11208.

Green, B. H., Copeland, J. R. M., Dewey, M. E., Sharma, V., Saunders, P. A., Davidson, I. A., … McWilliam, C. (1992). Risk factors for depression in elderly people: A prospective study. *Acta Psychiatrica Scandinavica*, *86*(3), 213–217.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43–49.

Hawton, K., & van Heeringen, K. (2009). Suicide. *Lancet*, *373*(9672), 1372–1381.

Homan, C. M., Johar, R., Liu, T., Lytle, M., Silenzio, V., & Alm, C.O. (2014). 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* (pp. 107–117)*.* Baltimore, MD: Association for Computational Linguistics.

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd edition). New York: Guilford Press (pp. 102–138).

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

Levi-Belz, Y., Gvion, Y., & Apter, A. (2019). Editorial: The psychology of suicide: From research understandings to intervention and treatment. *Frontiers in Psychiatry*, *10*, 214.

McCredie, M. N., & Morey, L. C. (2018). Who are the turkers? A characterization of MTurk workers using the personality assessment inventory. *Assessment*, 1073191118760709.

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy*, *28*(6), 487–495.

Mogotsi, I. C., Manning, C. D., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. *Information Retrieval*, *13*, 192–195.

Mundt, J. C., Greist, J. H., Gelenberg, A. J., Katzelnick, D. J., Jefferson, J. W., & Modell, J. G. (2010). Feasibility and validation of a computer-automated Columbia-Suicide Severity Rating Scale using interactive voice response technology. *Journal of Psychiatric Research*, *44*(16), 1224–1228.

Niederhoffer, K., Hollingshead, K., Resnik, P., Resnik, R., & Loveys, K. (2019). *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psycholo*gy.Stroudsberg, PA: Association for Computational Linguistics.

Nolen-Hoeksema, S., & Morrow, J. A. (1991). Prospective study of depression and posttraumatic stress symptoms after a natural disaster: The 1989 Loma Prieta earthquake. *Journal of Personality and Social Psychology*, *61*(1), 115–121.

Nolen-Hoeksema, S., & Watkins, E. R. (2011). A heuristic for developing transdiagnostic models of psychopathology: Explaining multifinality and divergent trajectories. *Perspectives on Psychological Science*, *6*(6), 589–609.

Ophir, Y., Asterhan, C. S. C., & Schwarz, B. B. (2017). Unfolding the notes from the walls: Adolescents’ depression manifestations on Facebook. *Computers in Human Behavior*, *72*, 96–107.

Ophir, Y., Asterhan, C. S. C., & Schwarz, B. B. (2019). The digital footprints of adolescent depression, social rejection and victimization of bullying on Facebook. *Computers in Human Behavior*, *91*, 62–71.

Ophir, Y., Sisso, L., Asterhan, C. S. C, Tikochinski, R., & Reichart, R. (2019, October 2). The turker blues: Hidden factors behind increased depression rates among Amazon’s mechanical turkers. *Clinical Psychological Science*. https://doi.org/10.1177/2167702619865973.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP) (pp. 1532–1543).

Peters, M. E., Neumann, M., Lyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, Z. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., … Mann, J. J. (2011). The Columbia–Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry*, *168*(12), 1266–1277.

Prims, J. P., Sisso, I., & Bai, H. (2018). Suspicious IP online flagging tool. Retrieved on October 28, 2019, from https://itaysisso.shinyapps.io/Bots.

Rammstedt, B., & John, O. P. (2007). 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*, *41*(1), 203–212.

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S.,  Danforth, C. M., & Langer, E. L. (2016, August 27). Forecasting the onset and course of mental illness with Twitter data. arXiv:1608.07740.

Ruder, S. (2017, June 15). An overview of multi-task learning in deep neural networks. *arXiv:1706.05098*.

Russell, D. W. (1996). UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment*, *66*(1), 20–40.

Salgado, J. F. (2018). 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*, *10*(1), 35–47.

Sartorius, N., Üstün, T. B., Lecrubier, Y., & Wittchen, H. U. (1996). Depression comorbid with anxiety: Results from the WHO study on psychological disorders in primary health care. *British Journal of Psychiatry*, *168*(S30), 38–43.

Sawhney, R., Manchanda, P., Singh, R., & Aggarwal, S. (2018). A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, student research workshop* (pp. 91–98).

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, *166*(10), 1092–1097.

Stack, S. (1983). The effect of religious commitment on suicide: A cross-national analysis. *Journal of Health and Social Behavior*, *24*(4), 362–374.

Statista. (2019). Most famous social network sites worldwide as of July 2019, ranked by number of active users (in millions). Retrieved on October 28, 2019, from http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users.

VanderWeele, T. J., Li, S., Tsai, A. C., & Kawachi, I. Association between religious service attendance and lower suicide rates among US women. *JAMA Psychiatry*, *73*, 845-851.

Viguera, A. C. , Milano, N., Laurel, R., Thompson, N. R., Griffith, S. D., Baldessarini, R. J., & Katzan, I. L. (2015). Comparison of electronic screening for suicidal risk with the Patient Health Questionnaire Item 9 and the Columbia Suicide Severity Rating Scale in an outpatient psychiatric clinic. *Psychosomatics*, *56*(5), 460–469.

Weber, A. N., Michail, M., Thompson, A., & Fiedorowicz, J. G. (2017). Psychiatric emergencies: Assessing and managing suicidal ideation. *The Medical Clinics of North America*, *101*(3), 553–571.

Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). 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* (pp. 24–33).

**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.

**Figure 3.** Illustration of the hierarchical “pyramid” of risk factors for suicide.

Note: The bottom of the proposed pyramid consists of the big five personality traits (i.e., openness; conscientious; extraversion; agreeableness; and neuroticism). The middle layers consist of the psychosocial risk factors (i.e., depressive rumination, worries, loneliness, and low satisfaction with life) and the psychiatric disorders (i.e., depression and anxiety), and the top layer consists of the predicted output, which is the two types of binary suicide variables (i.e., general and high suicide risk).

**Table 1.** Descriptive statistics and Correlations (N = 1,650 attentive users with at least one post).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Suicide | Depression | Anxiety | Brooding | Worry | SWL | Lonely | Open | Conscientious | Extravert | Agreeable | Neurotic |
| Means(SD) | 0.8(1.35) | 6.95(5.93) | 13.62(5.48) | 10.54(3.52) | 49.42(15.71) | 20.66(8.14) | 23.42(6.78) | 7.66(1.98) | 7.64(1.86) | 5.53(2.38) | 6.94(2.03) | 6.42(2.44) |
| Depression | .459\*\* |  |  |  |  |  |  |  |  |  |  |  |
| Anxiety | .381\*\* | .760\*\* |  |  |  |  |  |  |  |  |  |  |
| Brooding | .390\*\* | .624\*\* | .648\*\* |  |  |  |  |  |  |  |  |  |
| Worry | .331\*\* | .566\*\* | .714\*\* | .645\*\* |  |  |  |  |  |  |  |  |
| SWL | -.360\*\* | -.534\*\* | -.449\*\* | -.458\*\* | -.423\*\* |  |  |  |  |  |  |  |
| Lonely | .384\*\* | .599\*\* | .508\*\* | .548\*\* | .490\*\* | -.607\*\* |  |  |  |  |  |  |
| Open | .072\*\* | -.005 | .020 | .009 | -.006 | -.012 | -.059\* |  |  |  |  |  |
| Conscientious | -.185\*\* | -.341\*\* | -.226\*\* | -.293\*\* | -.224\*\* | .269\*\* | -.302\*\* | .103\*\* |  |  |  |  |
| Extravert | -.179\*\* | -.259\*\* | -.236\*\* | -.209\*\* | -.287\*\* | .273\*\* | -.395\*\* | .143\*\* | .153\*\* |  |  |  |
| Agreeable | -.209\*\* | -.273\*\* | -.301\*\* | -.234\*\* | -.280\*\* | .262\*\* | -.351\*\* | .040 | .113\*\* | .199\*\* |  |  |
| Neurotic | .315\*\* | .506\*\* | .628\*\* | .561\*\* | .779\*\* | -.393\*\* | .468\*\* | -.061\* | -.289\*\* | -.323\*\* | -.304\*\* |  |

Note: Suicide = the total score of the CSSRS; SWL = Satisfaction With Life scale. Notice that the current research addressed low satisfaction with life whereas the SWL is formulated in a positive manner (i.e., high satisfaction with life). This positive formulation explains the negative correlation between SWL and depression.

**Table 2.** 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.

**Table 3.** Hundred most frequent words that best distinguished between the four classes of prediction

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **True Positive** | **False Positive** | **True Negative** | **False Negative** |
| 1 | anymore | president | lord | fight |
| 2 | lose | husband | shall | ufb |
| 3 | kinda | season | blessed | office |
| 4 | probably | games | king | daddy |
| 5 | mad | gun | gift | fire |
| 6 | poor | strong | james | system |
| 7 | positive | wonder | christ | email |
| 8 | room | cat | church | experience |
| 9 | pretty | continue | missed | nobody |
| 10 | sitting | update | drink | shows |
| 11 | top | news | loving | ufd |
| 12 | cheese | damn | anybody | just |
| 13 | air | paid | god | america |
| 14 | eating | writing | monday | war |
| 15 | sick | none | wedding | none |
| 16 | okay | words | faith | questions |
| 17 | pain | talking | christmas | king |
| 18 | cry | change | jesus | ya |
| 19 | actually | ass | coffee | etc |
| 20 | instead | hot | prayer | pray |
| 21 | seriously | cold | however | like |
| 22 | easy | test | prayers | day |
| 23 | bit | watched | version | bus |
| 24 | reason | within | history | prayers |
| 25 | game | children | kill | one |
| 26 | clothes | fb | wishes | can |
| 27 | mother | asking | answer | happy |
| 28 | daughter | knew | state | know |
| 29 | hurt | grow | comment | get |
| 30 | worst | cancer | law | whats |
| 31 | bad | blood | important | will |
| 32 | account | shit | heaven | time |
| 33 | felt | red | mental | five |
| 34 | theyre | yesterday | father | texas |
| 35 | wow | dad | dr | sister |
| 36 | enjoy | holiday | ready | soul |
| 37 | lady | country | son | running |
| 38 | cut | near | pass | street |
| 39 | sleep | look | spirit | coming |
| 40 | fuck | loved | ufc | green |
| 41 | supposed | weeks | email | yeah |
| 42 | bed | looked | peace | american |
| 43 | pizza | funny | line | fine |
| 44 | quite | text | thanksgiving | along |
| 45 | gets | around | perfect | attention |
| 46 | guess | support | ufb | close |
| 47 | drive | relationship | war | human |
| 48 | door | course | save | died |
| 49 | thinking | putting | lol | company |
| 50 | surgery | wanted | together | problems |
| 51 | gonna | seems | dear | aint |
| 52 | literally | found | thank | business |
| 53 | thats | couple | just | prayer |
| 54 | sleeping | world | great | share |
| 55 | bitch | daily | sunday | area |
| 56 | cream | john | happy | listen |
| 57 | heart | looks | working | bit |
| 58 | wonderful | ask | fall | water |
| 59 | big | several | day | truly |
| 60 | arent | posts | cause | missing |
| 61 | might | moving | st | go |
| 62 | fucking | half | given | character |
| 63 | hospital | age | today | page |
| 64 | told | seeing | help | people |
| 65 | sad | company | men | store |
| 66 | doesnt | longer | love | love |
| 67 | couldnt | kept | choose | local |
| 68 | wall | months | friends | group |
| 69 | favorite | tv | brother | learned |
| 70 | taking | florida | others | retweeted |
| 71 | cleaning | high | holy | send |
| 72 | stupid | others | time | busy |
| 73 | nap | away | like | good |
| 74 | ugh | given | city | gonna |
| 75 | start | kids | everyone | song |
| 76 | entire | women | giving | now |
| 77 | brain | place | roll | anybody |
| 78 | car | hand | please | im |
| 79 | wear | sit | florida | park |
| 80 | times | weekend | mother | sorry |
| 81 | dinner | hey | copy | see |
| 82 | play | fear | question | pass |
| 83 | story | run | wife | gas |
| 84 | isnt | voice | know | birthday |
| 85 | calling | dark | friend | wishes |
| 86 | white | called | forget | asking |
| 87 | spent | spend | cleaning | end |
| 88 | mind | eyes | teacher | books |
| 89 | online | learn | get | lives |
| 90 | hopefully | light | group | miss |
| 91 | ice | history | pray | everyone |
| 92 | making | true | child | jesus |
| 93 | rest | saw | can | missed |
| 94 | feet | body | go | dark |
| 95 | order | heard | known | figure |
| 96 | cute | open | safe | wedding |
| 97 | understand | happened | one | entire |
| 98 | type | move | busy | today |
| 99 | sure | write | american | came |
| 100 | summer | due | lets | forget |

Note: This table presents the hundred most frequent words that best distinguished between the four classes of prediction (True Positive, True Negative, False Positive, and False Negative), using *Term Frequency Inverse Document Frequency (TF-IDF)*.

1. The 'general risk' group was chosen for the qualitative analysis because it is the larger group (compared with its sub-group of high risk individuals), therefore providing more textual content for the analysis. [↑](#footnote-ref-1)