Journal of the American Medical Informatics Association, 00(0), 2021, 1–10
doi: 10.1093/jamia/ocab031
Research and Applications

Research and Applications

Robust suicide risk assessment on social media via deep adversarial learning

Ramit Sawhney,[1]* Harshit Joshi,[2] Saumya Gandhi,[3] Di Jin,[4] and Rajiv Ratn Shah[1]

[1]MIDAS, IIIT Delhi, New Delhi, India, [2]Cluster Innovation Centre, University of Delhi, New Delhi, India, [3]Computer Science, Visvesvaraya National Institute of Technology, Nagpur, India, and [4]Computer Science and Artificial Intelligence Lab, Massachussetts Institute of Technology, Cambridge, Massachussetts USA

*Corresponding Author: Ramit Sawhney, BE, IIIT Delhi, Okhla Industrial Estate, Phase III, New Delhi, Delhi 110020 India (ramits@iiitd.ac.in)

Received 7 November 2020; Revised 6 January 2021; Editorial Decision 28 January 2021; Accepted 22 March 2021

ABSTRACT

Objective: The prevalence of social media for sharing personal thoughts makes it a viable platform for the assessment of suicide risk. However, deep learning models are not able to capture the diverse nature of linguistic choices and temporal patterns that can be exhibited by a suicidal user on social media and end up overfitting on specific cues that are not generally applicable. We propose Adversarial Suicide assessment Hierarchical Attention (ASHA), a hierarchical attention model that employs adversarial learning for improving the generalization ability of the model.

Material and Methods: We assess the suicide risk of a social media user across 5 levels of increasing severity of risk. ASHA leverages a transformer-based architecture to learn the semantic nature of social media posts and a temporal attention-based long short-term memory architecture for the sequential modeling of a user's historical posts. We dynamically generate adversarial examples by adding perturbations to actual examples that can simulate the stochasticity in historical posts, thereby making the model robust.

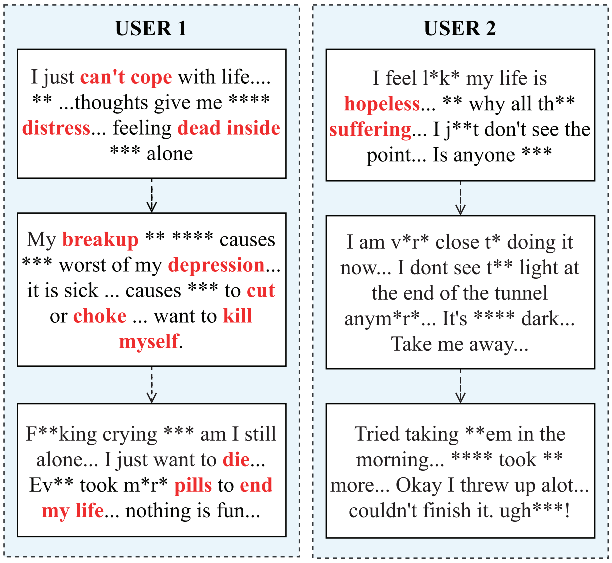
Results: Through extensive experiments, we establish the face-value of ASHA and show that it significantly outperforms existing baselines, with the F1 score of 64%. This is a 2% and a 4% increase over the ContextBERT and ContextCNN baselines, respectively. Finally, we discuss the practical applicability and ethical aspects of our work pertaining to ASHA, as a human-in-the-loop framework.

Discussion and Conclusions: Adversarial samples can be helpful in capturing the diverse nature of suicidal ideation. Through ASHA, we hope to form a component in a larger human-in-the-loop infrastructure for suicide risk assessment on social media.

Key words: machine learning, social media, suicidal ideation, adversarial learning, ordinal regression

INTRODUCTION

Suicide is a global phenomenon claiming the lives of nearly 800 000 people every year.[1] Despite the success of psychoclinical methods, such as the Suicide Probability Scale[2] and Suicide Ideation Questionnaire,[3,4] the suicide rate in the US has risen by 35% in the last 20 years.[5] While these methods are professional,[6] they have limited efficacy and may even impact participants negatively.[7] Unfortunately, 80% of patients do not undergo psychiatric treatment, and about 60% of those who died of suicide denied having any suicidal thoughts to mental health practitioners.[8] Contrastingly, social media's ubiquity presents a platform for individuals to express their feelings, instead of traditional, formal clinical settings,[9] with 8 out of 10 people disclosing their suicidal thoughts and plans.[10] Recently, natural language processing (NLP) methods using psycholinguistic lexicons[11–14] and textual features[15] have shown promise for suicide risk assessment, but analyzing individual posts alone is often insufficient to ascertain suicide risk. Analyzing a user's mental state through the emotions expressed overtime via their longitudinal social media posts in an automated manner can further contextualize potential suicidal risk. Although these classification methods show promise, it remains challenging to deploy them into real-world applications for multiple reasons.[16] First, existing methods treat suicide risk assessment as a coarse-grain binary classification task[11,15,17] that can lead to "*artificial notions of risk*."[18] Such a simplification of s could flag several users when deployed, likely exceeding the limited capacity of a resource-limited mental health ecosystem.[19,20] Hence, it is essential to perform a fine-grained assessment of risk to support expert-based intervention priority decisions. Another limitation of existing approaches is they treat all the risk levels equally, ignoring the inherent ordinal nature between risk levels. Models are not penalized relative to how far a prediction is from the actual risk level and consider wrongly predicting a high-risk user as no-risk or moderate-risk equally wrong. Lastly, existing suicide risk assessment methods are trained using relatively small datasets,[21,22] likely due to the paucity of expensive high-quality annotations by domain experts, such as clinicians and psychologists. They rely on a fixed set of suicide-related lexicons to develop a training set that may not encompass the diversity in expressing suicidal ideation, leading to linguistic and demographic biases in these datasets. Zhou and Mohit showed that NLP models tend to overfit on lexical biases present in a dataset that cannot cover the varying linguistic patterns and forms of suicidal expression.[23] For instance, in Figure 1, we see that while User 1 explicitly expresses suicidal thoughts by frequently using phrases found in the suicide-related lexicon (red), User 2 has a much more subtle form of expressing suicidal intent. NLP models tend to exploit such lexical and demographic biases without developing a "true" language understanding.[24,25] Such models' performance degrades drastically during deployment as they are unable to generalize to a broader spectrum of users.[23] To tackle this, we leverage adversarial training that has shown immense success in improving model generalization in image processing[26,27] and NLP.[28,29] Specifically, we perturb the joint vector representation of linguistic features and temporal patterns dynamically to generate additional adversarial examples for training.



I just **can't cope** with life.... ** ...thoughts give me **** **distress**... feeling **dead inside** *** alone

My **breakup** ** **** causes *** worst of my **depression**... it is sick ... causes *** to **cut** or **choke** ... want to **kill myself**.

F**king crying *** am I still alone... I just want to **die**... Ev** took m*r* **pills** to **end my life**... nothing is fun...

I feel l*k* my life is **hopeless**... ** why all th** **suffering**... I j**t don't see the point... Is anyone ***

I am v*r* close t* doing it now... I dont see t** light at the end of the tunnel anym*r*... It's **** dark... Take me away...

Tried taking **em in the morning... **** took ** more... Okay I threw up alot... couldn't finish it. ugh***!

Figure 1. We study the user history of 2 highly suicidal individuals on mental health sub-Reddits. Red font implies the use of words that were present in the suicide-related lexicon used for scraping posts for the dataset. Both users express similar feelings of hopelessness and distress in their situation and describe similar suicidal attempts. User 1 is very direct about his/her thoughts and frequently uses words found in the lexicon list. On the other hand, User 2 has a much more subtle expression with the use of phrases like "light at…tunnel," which are uncommon. A learning model may tend to overfit on commonly seen lexical choices as made by User 1 and possibly miss the subtle cues offered in posts made by User 2. All examples in this article have been paraphrased and obfuscated for user privacy.

Contributions: We reformulate suicide risk assessment as an *ordinal regression* problem, where *not all wrong classes are equally wrong*, by utilizing the inherent ordinal nature between the risk levels (Problem formulation). In this ordinal formulation, we map categorical risk levels to soft probability distributions to learn the natural interclass relationships between suicide risk levels. We propose to employ adversarial training to reduce the impact of dataset biases and data sparsity on predictive models' generalization ability. To this end, we present the Adversarial Suicide assessment Hierarchical Attention (ASHA) model, employing adversarial training on a hierarchical attention model. ASHA leverages a transformer-based architecture to learn the semantic nature of individual posts (Post embedding) and a temporal attention-based bidirectional long short-term memory (Bi-LSTM) architecture for the sequential modeling of the user history (User context modeling). We create adversarial training examples by perturbing the user history (Adversarial training), thereby increasing the model's robustness. Through a series of quantitative experiments and a qualitative analysis (RESULTS) on a Reddit dataset (Data), we analyze adversarial learning's impact on the model's robustness. Finally, we discuss the ethical considerations of our work and its practical applicability (Discussion).

MATERIALS AND METHODS

Problem formulation

We aim to assess the suicidal risk of a social media user $u_i \in U = \{u_1, u_2, \ldots, u_N\}$ by analyzing their historical posts $P_i = [\mathrm{p}^{\mathrm{i}}_1, \mathrm{p}^{\mathrm{i}}_2, \ldots, \mathrm{p}^{\mathrm{i}}_{\mathrm{Ti}}]$. P_i is arranged in a chronological manner, where $p^{\mathrm{i}}_{\mathrm{t}}$ is the t$^{\mathrm{th}}$ post made by the user with $p^i_{T_i}$ being the most recent post. We assess risk based on an adaptation of the Columbia Suicide Severity Rating scale (CSSRS)[30] for social media platforms.[21] As per the adapted CSSR scale, a user is assigned 1 of the 5 categories of increasing risk: Supportive (SU) < Indicator (IN) < Ideation (ID) < Behavior (BR) < Attempt (AT). We utilize the relative ordering present between the categories of risk by treating risk assessment as an *ordinal regression* problem. Such a formulation allows greater emphasis on the high-risk classes, in contrast to treating each risk category equally.

Our broader objective is to create a *robust* model that can perform a fine-grained assessment of risk for a wide spectrum of users. To this end, we present ASHA, a data-driven approach to aid the relevant healthcare stakeholders in the prioritization of at-risk users. ASHA is illustrated in Figure 2. Given the input of historic user posts, the post embedding layer encodes each post into a post vector e^i_t (Post Embedding). These post vectors are encoded by a BiLSTM. The temporal attention layer assigns an attention value to each post and calculates the weighted mean of these encoded post vectors to represent the overall sequential user context information a_i (User context modeling). Finally, we create adversarial examples by per-

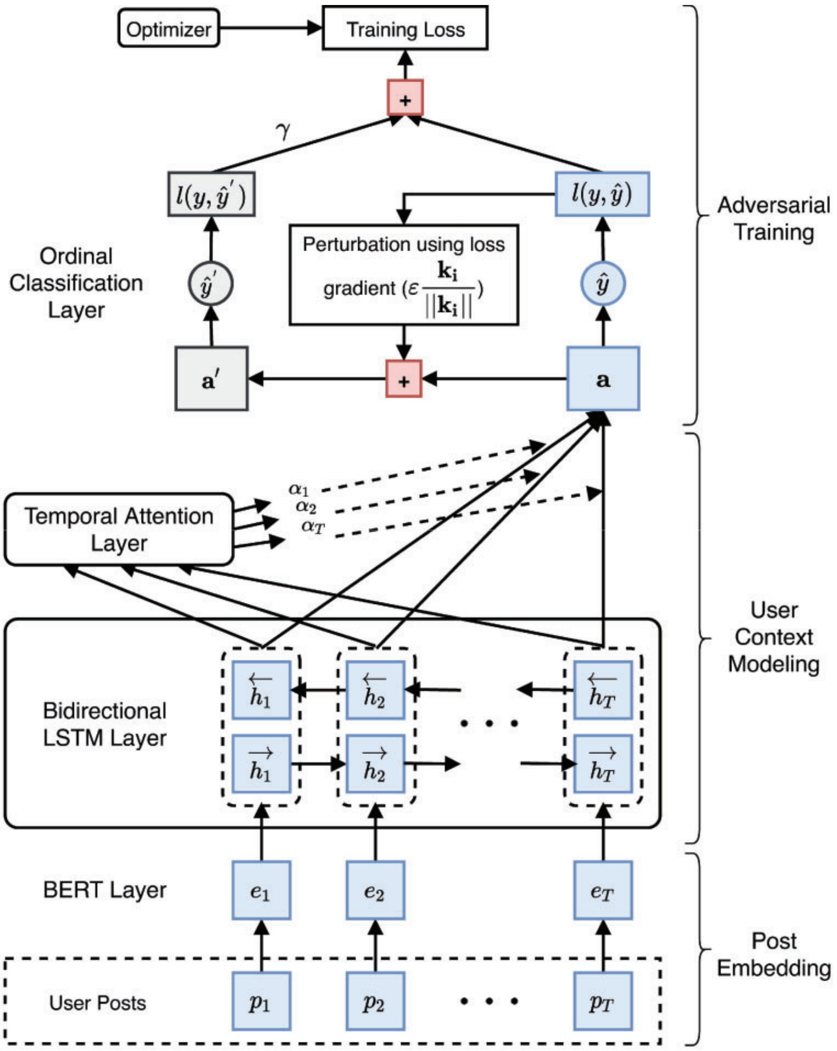


Figure 2. An illustration of adversarial suicide assessment hierarchical attention model.

turbing the final vector representation of the user context (Adversarial training).

Post embedding

Owing to the success of transfer learning and pretraining of language models in NLP and suicide risk assessment on social media,[9] we utilize the base version of bidirectional encoder representations from transformers (BERT)[31] model to derive comprehensive post-level embeddings to gain insights regarding a user's mental health.[12] We tokenize each historical post and add `[CLS]` token at the beginning of each post. We use the hidden state corresponding to this `[CLS]` token as the aggregate representation of the post. We encode each post p^i_t as: $e^i_t = \text{BERT}(p^i_t)$.

User context modeling

We now encode the historic posts of a user sequentially.

Bidirectional LSTM Layer: The buildup to suicide ideation can occur weeks, months, or even years before its onset.[32] The ability of LSTMs to capture long-term dependencies and extract temporal patterns[33] makes them suitable for processing the posts of a user sequentially. (Note: We acknowledge that LSTMs implicitly assume inputs to be equally spaced in time. A time-aware LSTM can replace the LSTM in ASHA without any other changes, if timestamps are available.). We utilize BiLSTM to capture the past and future context of a post. The encoding of the t^{th} post $\mathbf{e}_t{}^i$ is transformed into a contextualized representation h^i_t by concatenating the left-to-right and right-to-left hidden state vectors of the BiLSTM. Formally,

$$\vec{h^i_t} = LSTM(e^i_t, \vec{h^i_{t-1}}), \overleftarrow{h^i_t} = LSTM(e^i_t, \overleftarrow{h^i_{t-1}})$$

$$\mathbf{h}^i_t = [\vec{h^i_t}, \overleftarrow{h^i_t}]$$

The BiLSTM layer maps historical post encodings $[\mathbf{e}^i_1, \mathbf{e}^i_2, \ldots, \mathbf{e}^i_{Ti}]$ into contextual representations $\left[\mathbf{h}^i_1, \mathbf{h}^i_2, \ldots, \mathbf{h}^i_{Ti}\right] \in R^{H\times Ti}$ where H is the dimension of a hidden state vector and T_i is the total number of posts.

Temporal Attention: The suicidal risk associated with a user can vary over time, and the user's posts may be indicative of the varying suicide ideation associated with each post.[34,35] To this end, we adopt an attention mechanism over the representations of posts by a user over time. This mechanism applies adaptive weights for the contextual representation of each post $(\mathbf{h}^i_t)$, rewards posts with indicative markers for suicide risk, and aggregates them as:

$$\boldsymbol{a}_i = \sum_{t=1}^{T_i} \boldsymbol{\beta}_t^{i} \, \mathbf{h}_\mathrm{t}^\mathrm{i}, \;\; \boldsymbol{\beta}_t^i = \frac{\exp(\tilde{\beta}_t^i)}{\sum_{t=1}^{T_i} \exp(\tilde{\beta}_t^i)} \tilde{\beta}_t^i = c_x^T \tanh(\boldsymbol{W}_x \mathbf{h}_\mathrm{t}^\mathrm{i} + \boldsymbol{b}_x \;)$$

where $\boldsymbol{W}_x \in R^{E \times H}$, $\boldsymbol{b}_x \in R^E and \; c_x \in R^E$ are trainable network parameters, E is the dimension of the attention layer and $\boldsymbol{a}_i$ is the contextual representation of a user's historical posts. We refer to $\boldsymbol{a}_i$ as the clean example.

Ordinal Regression: The final discriminative network is a fully connected layer which takes input $\boldsymbol{a}_i$ to output the final classification confidence values for all levels of suicide risk $\widehat{\boldsymbol{y}}_i = \boldsymbol{W}_y \boldsymbol{a}_i + \boldsymbol{b}_y$ where $\boldsymbol{W}_y \in R^{H \times \omega}$ *and* $\boldsymbol{b}_y$ are trainable network parameters and ω denotes the number of classes.

To preserve the natural ordering of the severity of risk classes, we train the model by minimizing an ordinal regression loss.[36] Instead of using a one-hot vector representation of the ground truth, we use a soft encoded vector (probability distribution) representation and the classification score $\widehat{\boldsymbol{y}}$ to compute cross-entropy loss for optimization. Let $Y = \{SU = 0, \; IN = 1, \; ID = 2, \; BR = 3, \; AT = 4\} = \{r_i\}_{i=0}^4$ represent the 5 ordinal risk levels in the increasing order of severity. For a particular true risk level $r_t \in Y$, we compute soft labels as probability distributions $\boldsymbol{y} = [y_0, y_1, \ldots, \; y_4]$ of ground truth labels. The probability y_i of each risk level r_i is calculated as:

$$y_i = \frac{e^{-\varnothing(r_t, r_i)}}{\sum_{k=1}^{\omega} e^{-\varnothing(r_t, r_k)}}$$

where $\varnothing(r_t, r_i)$ is a precalculated interclass cost that penalizes how far the true risk level r_t is from a risk level $r_i \in Y$ and ω is the total number of risk levels (5). We select $\varnothing$ such that predicting a user labeled "Actual Attempt (AT)" (high-risk category) as Supportive (SU) should be strongly penalized as opposed to predicting Suicidal Behavior (BR), which is closer to AT on C-SSRS. We choose $\varnothing(r_t, \; r_i) = |r_t - r_i|$. After computing the probability distribution $\boldsymbol{y}_i$ for the ground-truth label, we finally calculate the cross-entropy loss for a sample using the classification confidence $\widehat{\boldsymbol{y}}_i$ as:

$$L_{ord}(\boldsymbol{y}_i, \widehat{\boldsymbol{y}}_i) = \sum_{j=1}^{\omega} y_{ij} \log(\widehat{y}_{ij})$$

where y_{ij} and $\widehat{y}_{ij}$ are the ground truth encoding and the confidence value, respectively, for class j and user $\boldsymbol{i}$. ω is the total number of risk levels (5). The final prediction is the risk level with the highest confidence score $\widehat{\boldsymbol{y}}_i$. We refer to the model that minimizes the objective function $L_{ord}(\boldsymbol{y}_i, \widehat{\boldsymbol{y}}_i)$ as **SHA** (Suicide-assessment Hierarchical Attention).

Adversarial training

The developmental trajectories of suicide-related outcomes can vary across different ages, ethnicities, and demographics.[37,38] Hence, optimizing the neural classification model by minimizing $L_{ord}(\boldsymbol{y}_i, \widehat{\boldsymbol{y}}_i)$ is not suitable as this form of training assumes the inputs to be static in nature. As such, the model may overfit on specific lexical markers of suicide ideation and the demographic biases present in the dataset. To this end, we apply adversarial learning. The intuition behind adversarial learning[26,39,40] is to simulate samples by adding intentional perturbations to the existing clean examples. The perturbation added to the clean example should be in the direction that leads to the largest change in model prediction. We apply perturbation to the embedding space of the user history ($\boldsymbol{a}_i$) to simulate the adversarial examples. We calculate the perturbation based on the embeddings of a user history $\boldsymbol{a}_i$ (clean example) and generate an adversarial example $\boldsymbol{a}_i^{'}$ as follows:

$$\boldsymbol{a}_i^{'} = \; \boldsymbol{a}_i + \varepsilon \frac{\boldsymbol{k}_i}{||\boldsymbol{k}_i||}, \;\; \boldsymbol{k}_i = \frac{\partial l(\boldsymbol{y}_i, \widehat{\boldsymbol{y}}_i)}{\partial \boldsymbol{a}_i}$$

where ε is the scale of perturbation which is a tunable hyperparameter and $\frac{\boldsymbol{k}_i}{||\boldsymbol{k}_i||}$ is a unit vector in the direction that leads to the largest change in model prediction. When ε is zero, no perturbation is added, and the model reduces to SHA. As we increase ε, the added perturbation increases. The value of ε needs to be adjusted during training to ensure that the added perturbation does not have an adverse effect on the model performance. We assign the same class to the adversarial example as that of its corresponding clean example. Training the model on clean as well as adversarial examples allows it to capture the stochastic and diverse nature of inputs. By incorporating the adversarial example, we finally optimize the model based on the joint objective function defined as:

$$L = \; \sum_{i=1}^{n} L_{ord}(\boldsymbol{y}_i, \widehat{\boldsymbol{y}}_i) + \; \gamma \sum_{i=1}^{n} L_{ord}(\boldsymbol{y}_i, \widehat{\boldsymbol{y}_i^{'}})$$

where $\widehat{\boldsymbol{y}_i^{'}}$ is the classification confidence of the adversarial examples and γ is a control parameter for balancing loss between clean and adversarial examples which are tunable hyperparameters.

EXPERIMENTS

Data

We use the dataset released by Gaur et al[21] that consists of Reddit posts of 500 users across 9 mental health and suicide-related subReddits. Annotation was performed at a user level across 5 categories of increasing suicide risk. The first such category comprises of *Supportive* (SU) users, those who engage in discussion on suicide-related forums without using language indicative of being at-risk in the past or the present. The second category, *Suicide Indicator* (IN), refers to the users who use at-risk language but are not actively experiencing any general or acute symptoms. The *Suicidal Ideation* (ID) category is defined for users who express thoughts of suicide or preoccupations with risk factors. Ideation can escalate to *Suicidal Behavior* (BR), operationalized as the confession of active or historical self-harm. The most severe level of risk, an *Actual Attempt* (AT) is defined as any deliberate action that could have resulted in intentional death, be it a completed attempt or not. The class-wise distributions of users for SU, IN, ID, BR, and AT are 22%, 20%, 34%, 15%, and 9%, respectively. Annotation was performed by 4 practicing psychiatrists following guidelines as per the C-SSRS. We observe that the posts made by each user vary to a large extent. On average, the number of posts made by a user is 18.25 ± 27.45. The median number of posts is 8 and the interquartile range is 19.

Preprocessing and Data Split: We deidentified the dataset by performing named entity recognition and removing any identifiable information, such as e-mail addresses, URLs, and names. Next, we convert the text to lowercase, remove punctuation and accents, strip white spaces and remove stop words. Due to the token limit imposed by BERT, we trim the posts having more than 510 words (0.73% of the total posts). Following Gaur et al,[21] we perform a stratified 80:20 split.

Evaluation metrics

To better evaluate the model's performance, Bryan and Rudd[18] proposed an altered formulation of False Negative (FN) and False Positive (FP). They define FN as the ratio of the number of times predicted suicide severity risk level (r') is less than actual level (r^o) over the size of test data (N_T). FP is defined as the ratio of the number of times r' is greater than r^o. They termed precision as Graded Precision (GP) and recall as Graded Recall (GR). Ordinal Error (OE) was defined as the ratio of the number of samples where the difference between r^o and r' is greater than 1. Formally,

$$FN = \frac{\sum_{i=1}^{N_T} I(r_i^o > r_i')}{N_T}, \quad FP = \frac{\sum_{i=1}^{N_T} I(r_i^o < r_i')}{N_T}$$

$$OE = \frac{\sum_{i=1}^{N_T} I\Big(\Delta(r_i^o, r_i') > 1\Big)}{N_T}$$

where $\Delta(r_i^o, r_i')$ is the difference between r_i^o and r_i'. r_i^o and r_i' are the actual and predicted response for i^{th} user.

Experimental setup

We select hyperparameters based on the highest F1 obtained through cross-validation for each model. We use grid-search to explore: number of features in hidden state $\mathbf{h}_t^i \in \{8,64,128,256,512\}$, number of LSTM layers $n \in \{1,2,5\}$, dropout $d \in \{0.3,0.4,0.5,0.6,0.7\}$, $\varepsilon \in \{0.05,0.1,0.15,0.2,0.25\}$, and $\gamma \in \{0.1,0.3,0.6,0.9\}$. The optimal hyperparameters found were $\mathbf{h}_t^i = 256, n = 1, d = 0.4, \varepsilon = 0.1$ and $\gamma = 0.3$.

RESULTS

Firstly, we evaluate the performance of ASHA against the baselines in Table 1. We further establish the utility of adversarial learning through the analysis of training-validation loss (and classification confidence in comparison to SHA. We perform an ablation study to determine the impact of various components of ASHA on its performance. Finally, we perform a manual label assessment, where we analyze the posts of 4 users and reason the corresponding model predictions by SHA and ASHA.

Performance comparison

From Table 1, we see that deep learning approaches outperform the handcrafted feature-based models, such as the SVM, Random Forest, and MLP. This can be attributed to the ability of contextual models to learn better representations of a user's mental state over time. Among contextual models, we find that sequential context models, such as the ContextBERT, SHA and ASHA, perform better than the ContextCNN model. We believe this is because sequential models can learn the temporal patterns that could be indicative of suicidal risk, in contrast to the ContextCNN's "bag of posts" approach. We also observe that the performance of ASHA exceeds that of ContextBERT across all metrics.

From Figure 3, we see that nonadversarial models overfit on predicting specific risk levels. For example, we see that C-BERT is highly inclined to predict ID for users of all categories, especially high-risk categories. On the other hand, predictions made by ASHA are spread more evenly across the diagonals, implying that it is able to predict varying risk levels correctly.

Training and validation loss analysis

While training the model, we evaluate it on the held-out validation set after each epoch of training to check the performance of the model on previously unseen data. In Figure 4, we see that the training loss for ASHA as well as SHA steadily decreases with more epochs and converges to zero toward the end of the training. However, we observe that the validation loss for SHA keeps fluctuating around the initial loss values in contrast to ASHA, whose validation loss gradually decreases with more epochs. The patterns observed in SHA are generally indicative of overfitting on the training examples learning specific temporal patterns found in the training set. Whereas, ASHA accounts for the stochastic nature of user histories, preventing the model from overfitting on temporal patterns present in the dataset.

Analysis of classification confidence

In Figure 5, we show how ASHA's classification confidence broadens on adding adversarial training. A classification confidence score is the softmax probability of predicted labels with which a classifier outputs the predicted class for each sample. Firstly, we notice that the median of confidence distribution for ASHA is higher in comparison to its nonadversarial variant (SHA) implying that it is, in general, more confident of its predictions. This can be attributed to ASHA learning across a wider range of examples due to adversarial training which helps in classifying previously unseen samples with more confidence. We also observe that the interquartile range as well as the maxima–minima range is much larger for ASHA. This is a result of the adversarial training pushing the classification decision boundary far from clean examples, which enhances ASHA's generalization.

Ablation study

We perform an ablation study to assess the impact of ordinal loss, sequential modeling, and adversarial training on model performance, as shown in Table 2. We train an average pooling model that obtains user representations by averaging all contextualized post representations e_i^t. The average pooling model and the ContextCNN (BERT) both utilize BERT embeddings and aggregate user posts without utilizing their relative ordering, but the average pooling model shows an improved recall, which can be due to the ordinal loss component. We find that the sequential SHA-O model outperforms both the nonsequential models. This can suggest the presence of temporal patterns in the user posting behavior, which can be indicative of suicidal risk. We also notice that adding the ordinal component (SHA) or the adversarial component (ASHA-O) to SHA-O improves the model performance, highlighting the individual utility of both the components. Finally, the combined effect of adversarial learning and ordinal loss can be seen in ASHA, which outperforms all previous models in the ablation study.

Manual label assessment

In Figure 6, we qualitatively analyze 4 users and reason the corresponding model predictions. User A frequently uses phrases common amongst at-risk users on social media, such as "kill myself," or "hang myself," which are explicitly suggestive of suicidal tendencies. This direct manner of expression allows both the models to correctly predict the highest level of risk for the user. Although User B is equally suicidal, the linguistic style greatly differs from User A. User B expresses suicidality in a reserved manner, which does not overlap with the suicide-related lexicon. Although words like "ropeee" (rope) and "hearrbreak" (heartbreak) are part of the fixed suicide-related lexicon, they did not match since these words were misspelled by the user, which is a common occurrence on social media. SHA incorrectly classifies User B as Supportive, possibly due to the presence of positive phrases such as "very nice" and "is helping"

Table 1. Classification is performed across all 5 classes. We report the mean of results obtained over 50 runs. * and + indicates that the result is statistically significant ($P < .05$ for Wilcoxon Signed Rank Test) compared to ContextCNN and ContextBERT, respectively. **Bold** denotes best performance across metric

Input Features	Model	Description	Training Loss	GP↑	GR↑	F1↑	OE↓
Language based features (part of speech tags, linguistic inquiry and word count)	SVM+RBF[21]	Language vector fed to SVM+RBF kernel with kernel parameter σ=0.24, cost parameter c = 5	Hinge	0.53	0.51	0.52	0.12
	Random forest[21]	RF with 500 trees and 9 predictor variables per tree node is trained using user's language vector	Gini Impurity	0.68	0.49	0.57	0.19
	MLP[21]	MLP with 2 hidden layers (64 dim)	Log Loss	0.45	0.59	0.51	0.15
GloVe embeddings	ContextCNN[21]	GloVe word embeddings from all posts of a user are concatenated and fed to the ContextCNN	Cross Entropy	0.69	0.52	0.59	0.10
BERT embeddings	ContextBERT[41]	Post-level BERT embeddings are fed to GRU cell	Cross Entropy	0.64	0.58*	0.60	0.10
	SHA (Our model)	Post embeddings (BERT) are fed to a BiLSTM + temporal attention layer with ordinal soft labels for implementing ordinal loss	Ordinal	0.66+	0.59*	0.62*+	0.10
	ASHA (Our model)	Adversarial samples generated and used for training ASHA to make it robust	Adversarial +Ordinal	**0.69**+	**0.59***	**0.64***+	**0.10**

Abbreviations: ASHA, Adversarial Suicide assessment Hierarchical Attention; MLP, multiple layer perceptron; RF, random forest; SHA, Suicide-assessment Hierarchical Attention; SVM, support vector machine; RBF, Radial Basis Function.

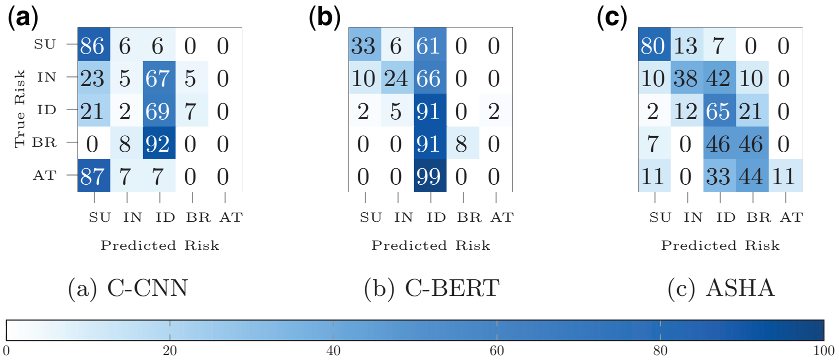
that are frequent among supportive users. In contrast, ASHA correctly assesses the risk, which could be due to an increased focus on the post's context rather than specific words or phrases. Users C and D tend to use phrases common among suicidal ideators, for example, "committed suicide" and "suicidal thoughts." User C belongs to a low-risk group, and SHA incorrectly perceives a higher risk than actual, due to possible overfitting on such phrases. However, ASHA precisely predicts the actual risk level, which can be attributed to the reduced impact of data sparsity for rarer classes, such as the indicative class [IN]. On the other hand, User D is a high-risk user. The overfitting tendencies of SHA prompts an accurate higher risk prediction than ASHA, which tries to regularize the dependency on lexicons, leading to an incorrect prediction. However, the ordinal component of ASHA aids in predicting a risk level close to the actual risk, which resembles a human's judgment of risk[42] and is crucial in practical settings.

DISCUSSION

Owing to the sensitive nature of the work presented in this article, we must discuss the ethical implications, biases involved, and the practical applicability of such models for suicidal risk assessment.

Ethical considerations: Due to a large number of personal thoughts and details embedded across social media posts made by a user manifest, its utility in understanding mental health stands in tension with concerns over user privacy. We work within the purview of acceptable privacy practices to avoid coercion and intrusive treatment. Our research involves the publicly available *Reddit* dataset introduced in Gaur et al.[21] Although *Reddit* is officially intended for anonymous posting,[43] we take further precautions by performing automatic deidentification of the dataset using named entity recognition.[19] All posts shown in Figures 1 and 6 have been paraphrased as per the *moderate disguise* scheme suggested by Bruckman.[44] Our work focuses on developing a neural model for screening of suicide risk and does not make any diagnostic claims related to suicide. The risk assessment made by our model is deeply personal and sensitive in nature and should be shared selectively to avoid misuse such as Samaritan's Radar.[45]

Bias: A fixed set of suicide related lexicons were used to annotate the data studied in this article, making it susceptible to demographic and medium-specific biases which can cause latent problems, especially when inferences are incorporated in real-life



(a) C-CNN (b) C-BERT (c) ASHA

Figure 3. Normalized (percentage) Confusion matrix for the median of the test results. A row represents the actual risk level, whereas a column represents the predicted risk level. For example, if we look at the element with True Risk AT and Predicted risk SU in (a), we can infer that the CCNN wrongly predicted 87% users in the Actual Attempt (AT) category as Supportive (SU) users. In the same way, the diagonal elements represent correct predictions.

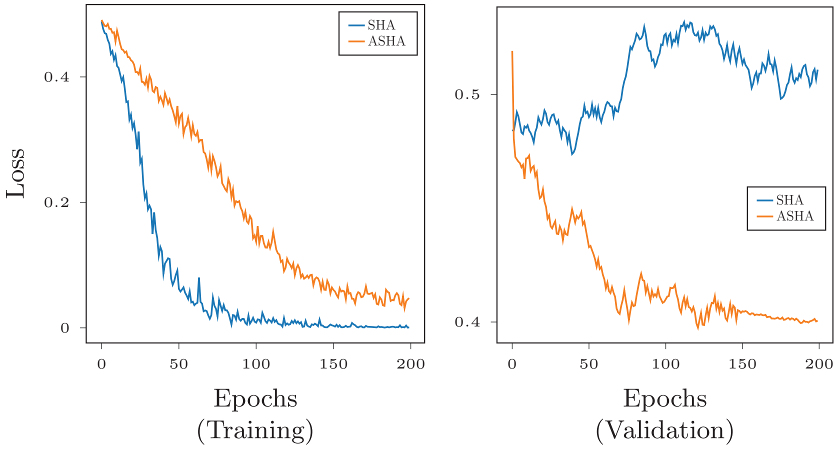


Figure 4. We analyze the training and validation loss during the training process over 200 epochs.

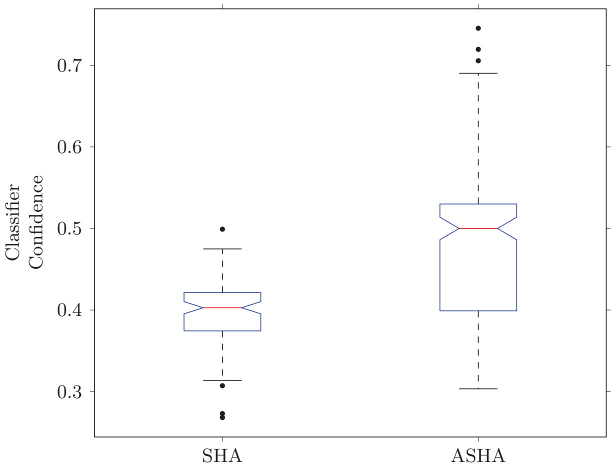


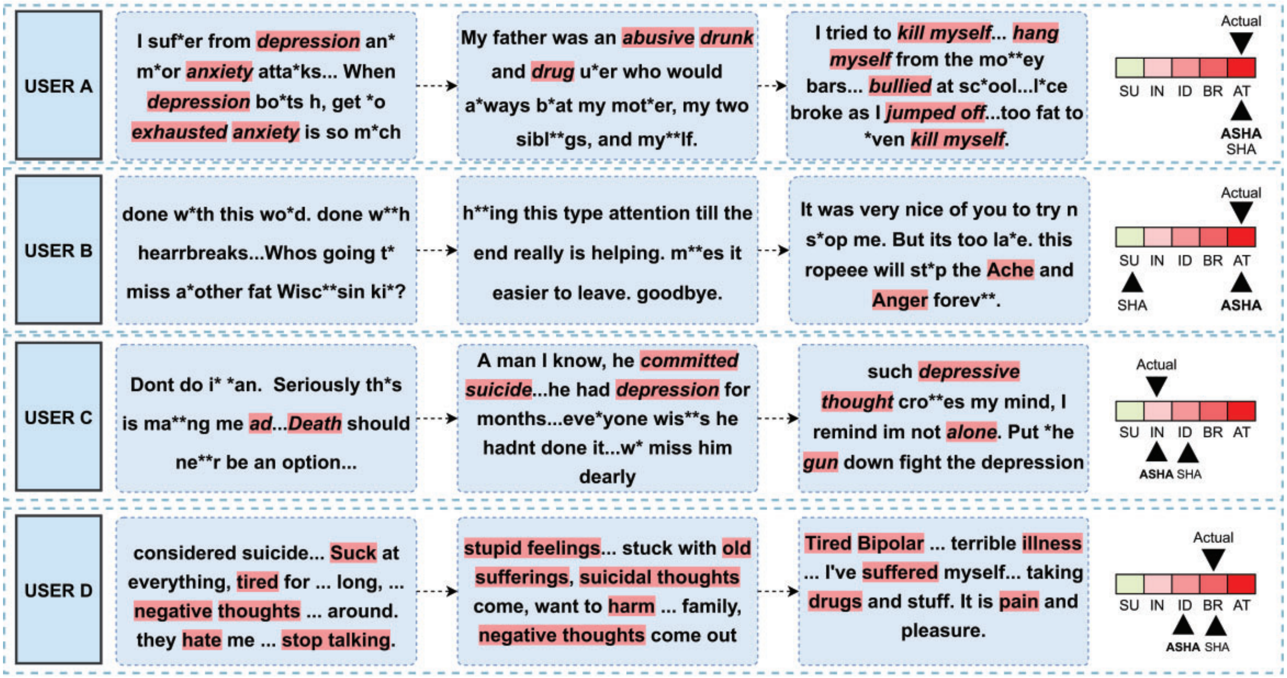
Figure 5. Classification confidence distribution for the predicted risk level of examples in the test set. The median, interquartile range and maxima–minima range is depicted by the red line, the blue outlined region, and the length of the dashed line, respectively.

situations. We attempt to reduce these existing biases in the dataset by supplying the model with adversarial examples during training. Although it is unclear to what extent online expressions of suicide ideation are comparable to suicidal risk as diagnosed by clinicians, recent studies do show a correlation of suicidal expression online with psychometrically assessed suicidal risk.[46,47] Annotation of the dataset was performed by practicing psychiatrists following the guidelines outlined in C-SSRS,[30] hence reducing the possibility of annotator specific biases in the dataset. However, we acknowledge that the study of suicidal risk is subjective in nature, and the interpretation of the analysis presented may vary across individuals. Although we believe that adversarial learning might help the model generalize across social media platforms, we acknowledge that the statistical patterns learned by ASHA may fail to generalize to different social networking websites and contexts that do not focus on mental health.[48] We also acknowledge that there exists a tradeoff between the inherent selection bias in the studied data and informed user consent.[49]

Practical Applicability and Future Work: We study the *Reddit* data in a purely observational and nonintrusive manner, without any intervention or alteration in the user experience. ASHA aims to provide a preliminary assessment of suicidal risk through ordinal regression that can aid in the prioritization of clinical resources.[20] It should form part of a human-centered mental health ecosystem involving psychiatrists, healthcare providers, internet-based Cognitive Behavioral Therapy (iCBT) staff, etc. Evaluating ASHA across social media platforms paves the way for our future work.

Table 2. We perform an ablation study by adding the ordinal, sequential, and adversarial modules to each subsequent model. SHA-O and ASHA-O represents the SHA and the ASHA models trained on Cross Entropy loss (nonordinal). We report the mean ± standard deviation across 50 runs. * indicates the result is statistically significant compared to ContextCNN ($P < .05$ for Wilcoxon Signed Rank Test). **Bold** denotes the best performance for a metric

Model	Ordinal loss	Sequential modeling	Adversarial training	Graded precision↑	Graded recall↑	F1↑	OE↓
ContextCNN (BERT)	×	×	×	0.64 ± 0.04	0.53 ± 0.04	0.58 ± 0.02	0.10 ± 0.04
Average pooling	✓	×	×	0.56 ± 0.04	0.59 ± 0.02	0.58 ± 0.02	0.12 ± 0.03
SHA-O	×	✓	×	0.64 ± 0.07	0.58* ±0.05	0.60 ± 0.05	0.10 ± 0.04
SHA	✓	✓	×	0.66 ± 0.06	0.59* ±0.04	0.62* ±0.04	0.10 ± 0.02
ASHA-O	×	✓	✓	0.65 ± 0.06	0.59* ±0.04	0.62* ±0.04	0.10 ± 0.02
ASHA	✓	✓	✓	**0.69** ± 0.05	**0.59*** ±0.04	**0.64*** ±0.04	**0.10** ± 0.02



I suf*er from depression an* m*or anxiety atta*ks... When depression bo*ts h, get *o exhausted anxiety is so m*ch

My father was an abusive drunk and drug u*er who would a*ways b*at my mot*er, my two sibl**gs, and my**lf.

I tried to kill myself... hang myself from the mo**ey bars... bullied at sc*ool...I*ce broke as I jumped off...too fat to *ven kill myself.

done w*th this wo*d. done w**h hearrbreaks...Whos going t* miss a*other fat Wisc**sin ki*?

h**ing this type attention till the end really is helping. m**es it easier to leave. goodbye.

It was very nice of you to try n s*op me. But its too la*e. this ropeee will st*p the Ache and Anger forev**.

Dont do i* *an. Seriously th*s is ma**ng me ad...Death should ne**r be an option...

A man I know, he committed suicide...he had depression for months...eve*yone wis**s he hadnt done it...w* miss him dearly

such depressive thought cro**es my mind, I remind im not alone. Put *he gun down fight the depression

considered suicide... Suck at everything, tired for ... long, ... negative thoughts ... around. they hate me ... stop talking.

stupid feelings... stuck with old sufferings, suicidal thoughts come, want to harm ... family, negative thoughts come out

Tired Bipolar ... terrible illness ... I've suffered myself... taking drugs and stuff. It is pain and pleasure.

Figure 6. We display the posts for 4 users from the test set. Posts are arranged chronologically from left to right. Alongside, we show the actual risk level of the user and the predictions made by SHA and ASHA. → depicts the presence of posts not displayed for brevity. Words/phrases in red indicate their overlap with the suicide-related lexicon.

CONCLUSION

In this work, building on the fine-grained assessment of suicidal risk on social media, we reformulate suicide risk assessment as an ordinal regression problem. We present ASHA, an adversarial hierarchical attention method to tackle the data sparsity, lexical biases, and stochastic nature of user history present in suicide risk assessment datasets, to build a robust model that can generalize in practical settings. We conducted extensive experiments to determine the validity of adversarial learning in lexical debiasing and accounting for the stochastic nature of user history. While our method aims to reduce the impact of dataset biases, the priority should be to curate bias-free datasets. However, some biases are inevitably present due to the nature of the task, and hence these methods can be helpful. Through ASHA, we hope to form a component in a larger human-in-the-loop infrastructure for suicide risk assessment on social media.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

RS contributed to the conception of the project. RS, HJ and SG contributed to the design of the project. HJ and SG contributed to the implementation of the experiments. All authors contributed to the interpretation of results. SG and HJ wrote the first draft of the article. RS and DJ critically reviewed the method descriptions and formulations. All authors provided critical feedback on and approved the final version of the article.

DATA AVAILABILITY

The data underlying this article are available at https://github.com/jpsain/Suicide-Severity.The source code for ASHA is available at https://github.com/midas-research/asha-jamia.

ETHICAL STATEMENT

This study was exempted from review from the institutional review board at IIIT Delhi, India.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. World Health Organization. "Preventing suicide: A global imperative." (2014).
2. Bagge C, Osman A. The suicide probability scale: Norms and factor structure. *Psychol Rep* 1998; 83 (2): 637–8.
3. Fu K-W, Liu KY, Yip PS. Predictive validity of the chinese version of the adult suicidal ideation questionnaire: psychometric properties and its short version. *Psychol Assess*2007; 19 (4): 422–9.
4. Hawton K, Witt KG, Salisbury TLT, *et al.* Psychosocial interventions following self-harm in adults: a systematic review and meta-analysis. *Lancet Psychiatry* 2016 Aug; 3 (8): 740–50.
5. Hedegaard H, Curtin SC, Warner M. Increase in suicide mortality in the United States, 1999–2018. *NCHS Data Brief* 2020; (362): 1–8.
6. Pestian JP, Sorter M, Connolly B, the STM Research Group, *et al.* A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. *Suicide Life Threat Behav* 2017; 47 (1): 112–21.
7. Harris KM, Ting-Ting Goh M. Is suicide assessment harmful to participants? Findings from a randomized controlled trial. *Int J Mental Health Nurs* 2017; 26 (2): 181–90.
8. McHugh CM, Corderoy A, Ryan CJ, Hickie IB, Large MM. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity, and positive predictive value [published correction appears in BJPsych Open]. *BJPsych Open* 2019 Mar; 5 (2): e18–24.
9. Coppersmith G, Leary R, Crutchley P, Fine A. Natural language processing of social media as screening for suicide risk. *Biomed Inform Insights* 2018; 10: 117822261879286.
10. Golden RN, Weiland C, Peterson F. *The Truth About Illness and Disease*. New York, NY: Facts on File of Infobase Publishing; 2009.
11. Sawhney R, Manchanda P, Singh R, *et al.* A computational approach to feature extraction for identification of suicidal ideation in tweets. In: Proceedings of ACL 2018, Student Research Workshop; July 2018: 91–8; Melbourne, Australia.
12. De Choudhury M, Gamon M, Counts S, *et al.* Predicting depression via social media. In: Proceedings of the International AAAI Conference on Web and Social Media*;* July 2013; 13: 1–10; Massachusetts, USA.
13. De Choudhury M, Kiciman E, Dredze M, *et al.* Discovering shifts to suicidal ideation from mental health content in social media. In: Proceedings of the 2016 CHI conference on human factors in computing systems; May 2016: 2098–110; California, USA.
14. Sawhney R, Manchanda P, Mathur P, *et al.* Exploring and learning suicidal ideation connotations on social media with deep learning. In: Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis; October 2018: 167–75; Brussels, Belgium.
15. Cao L, Zhang H, Feng L, *et al.* Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 2019: 1718–28; Hong Kong, China.
16. Chikersal P, Belgrave D, Doherty G, *et al.* Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In: Proceedings of the 2020 CHI conference on human factors in computing systems; April 2020: 1–16; Honolulu HI, USA.
17. Losada DE, Fabio C, Javier P. Overview of eRisk at CLEF 2019: early risk prediction on the internet (extended overview). In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2019;* August2019; Lugano, Switzerland.
18. Bryan CJ, Rudd MD. Advances in the assessment of suicide risk. *J Clin Psychol* 2006; 62 (2): 185–200.
19. Zirikly A, Resnik P, Uzuner O, *et al.* 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; June 2019: 24–33; Minneapolis, Minnesota.
20. Shing H-C, Resnik, P, Oard DW. A prioritization model for suicidality risk assessment. In: Proceedings of the 58th annual meeting of the association for computational linguistics; July 2020: 8124–37; Online.
21. Gaur M, Alambo A, Prakash Sain J, *et al.* Knowledge-aware assessment of severity of suicide risk for early intervention. In: The World Wide Web conference; May 2019: 514–25; San Francisco CA, USA.
22. Rohan , M, Prakhar Sinha P, Sawhney R, *et al.* SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media. In: Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: student research workshop; June 2019: 147–56; Minneapolis, Minnesota.
23. Zhou X, Mohit B. Towards robustifying NLI models against lexical dataset biases. *arXiv preprint arXiv:2005.04732* (2020).
24. Nie Y, Wang Y, Bansal M. Analyzing compositionality-sensitivity of NLI models. *AAAI* 2019; 33: 6867–74.
25. Wallace E, Rodriguez P, Feng S, *et al.* Trick me if you can: human-in-the-loop generation of adversarial examples for question answering. *Trans AssociationComput Linguist* 2019; 7: 387–401.
26. Goodfellow IJ, Jonathon S, Christian S. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
27. Kurakin A, Ian G, Samy B. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* (2016).
28. Jia R, Percy L. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328* (2017).
29. Caswell I, Onkur S, Allen N. Exploring adversarial learning on neural network models for text classification. https://web.stanford.edu/~anie/papers/CS224N_Final_Adversarial_Learning.pdf Accessed October 16, 2020.
30. Posner K, Brown GK, Stanley B, *et al.* The Columbia–Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Ajp* 2011; 168 (12): 1266–77.
31. Devlin J, Chang M, Lee K, *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
32. Overholser J. Predisposing factors in suicide attempts: life stressors. In A. Spirito A, Overholser JC, eds. *Evaluating and Treating Adolescent Suicide Attempters*. Washington, DC: Academic Press; 2003: 41–52.
33. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9 (8): 1735–80.
34. Nobles AL, Glenn JJ, Kowsari K, *et al.* Identification of imminent suicide risk among young adults using text messages. In: Proceedings of the 2018 CHI conference on human factors in computing systems; April2018; Montreal QC, Canada.
35. Glenn JJ, Nobles AL, Barnes LE, *et al.* Can text messages identify suicide risk in real time? a within-subjects pilot examination of temporally sensitive markers of suicide risk. *Clin Psychol Sci* 2020; 8 (4): 704–22.

36. Diaz R, Amit M. Soft labels for ordinal regression. In: Proceedings of the IEEE conference on computer vision and pattern recognition; June 2019; California, USA.
37. Czyz EK, King CA. Longitudinal trajectories of suicidal ideation and subsequent suicide attempts among adolescent inpatients. *J Clin Child Adolescent Psychol* 2015; 44 (1): 181–93.
38. Erausquin JT, McCoy TP, Bartlett R, *et al.* Trajectories of suicide ideation and attempts from early adolescence to mid-adulthood: associations with race/ethnicity. *J Youth Adolescence* 2019; 48 (9): 1796–805.
39. Tramèr F, Kurakin A, Papernot N, *et al.* Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204* (2017).
40. Feng F, Chen H, He X, *et al.* Enhancing stock movement prediction with adversarial training. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence; August 2019; Macao, China.
41. Matero M, Idani A, Son Y, *et al.* Suicide risk assessment with multi-level dual-context language and bert. In: Proceedings of the sixth workshop on computational linguistics and clinical psychology; June 2019; Minneapolis, Minnesota.
42. Ralf B, Ulrich G. Ordinal logistic regression in medical research. *J R Coll Physicians Lond* 1997; 31 (5): 546–51.
43. Gutman R. (2018). Reddit's case for anonymity on the internet.
44. Bruckman A. Studying the amateur artist: a perspective on disguising data collected in human subjects research on the Internet. *Ethics Inform Technol* 2002; 4 (3): 217–31.
45. Hsin H, Torous J, Roberts L. An adjuvant role for mobile health in psychiatry. *JAMA Psychiatry* 2016; 73 (2): 103–4.
46. Sueki H. The association of suicide-related Twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan. *J Affective Disord* 2015; 170: 155–60.
47. Berryman C, Ferguson CJ, Negy C. Social media use and mental health among young adults. *Psychiatr Q* 2018; 89 (2): 307–14.
48. Davis JL, Love TP. Generalizing from social media data: a formal theory approach. *Inform Commun Soc* 2019; 22 (5): 637–47.
49. Fiesler C, Proferes N. Participant perceptions of Twitter research ethics. *Social Media+ Society* 2018; 4 (1): 205630511876336.