Towards Ordinal Suicide Ideation Detection on Social Media

Ramit Sawhney
ramits@iiitd.ac.in
IIIT Delhi
India

Harshit Joshi
harshit113@ducic.ac.in
University of Delhi
India

Saumya Gandhi
gandhisaumya8@gmail.com
Visvesvaraya National Institute Of Technology
India

Rajiv Ratn Shah
rajivratn@iiitd.ac.in
IIIT Delhi
India

ABSTRACT

The rising ubiquity of social media presents a platform for individuals to express suicide ideation, instead of traditional, formal clinical settings. While neural methods for assessing suicide risk on social media have shown promise, a crippling limitation of existing solutions is that they ignore the inherent ordinal nature across fine-grain levels of suicide risk. To this end, we reformulate suicide risk assessment as an *Ordinal Regression* problem, over the Columbia-Suicide Severity Scale. We propose SISMO, a hierarchical attention model optimized to factor in the graded nature of increasing suicide risk levels, through soft probability distribution since not all wrong risk-levels are equally wrong. We establish the face value of SISMO for preliminary suicide risk assessment on real-world Reddit data annotated by clinical experts. We conclude by discussing the empirical, practical, and ethical considerations pertaining to SISMO in a larger picture, as a human-in-the-loop framework.[1]

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*; • **Human-centered computing** → **Social media**.

KEYWORDS

social media; suicide ideation; ordinal regression; reddit

ACM Reference Format:
Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021. Towards Ordinal Suicide Ideation Detection on Social Media. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441805

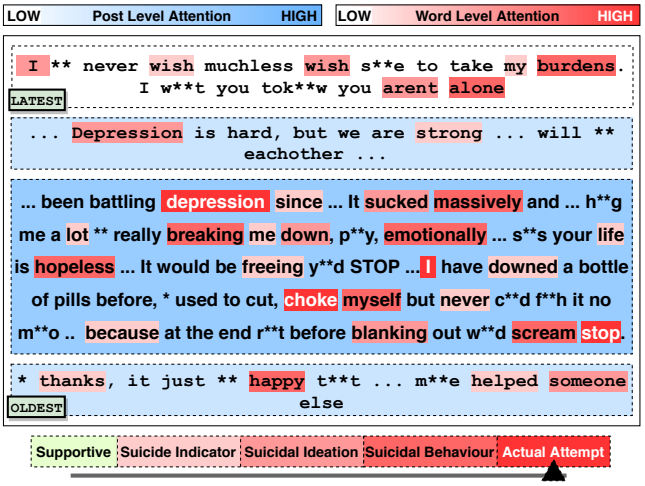


Figure 1: Example of a suicide risk assessment framework that can highlight possibly concerning posts (blue) and rank users across categories of increasing risk for subsequent analysis. The risk markers (red) highlight the foreground importance of information for the severity of detecting suicide ideation. The scale shows that the user is categorized in the "Actual Attempt" group by practicing psychiatrists under the Columbia-Suicide Severity Rating Scale. Such a framework can be used for preliminary screening in a larger clinical assessment infrastructure. *All examples in this paper have been paraphrased and obfuscated for user privacy.*

1 INTRODUCTION

Every 10.9 minutes, a person dies of suicide, and 25 times more people attempt suicide in the U.S. [16]. Suicide ranks as the second leading cause of death for 14-35 year-olds, and the fourth leading cause for people aged 35-64, with a rising rate of 35% since 1999 [25]. Extending appropriate clinical and psychological care to suicidal people relies on identifying those at-risk, and the degree of this suicide risk. Unfortunately, recent studies [20] suggest that the predictive power of existing clinical methods for assessing suicide risk has not advanced significantly from the past fifty years of research. Moreover, 80% of patients do not undergo clinical treatment, and about 60% of those who died of suicide denied having any suicidal thoughts to mental health practitioners [35]. In contrast, people exhibiting suicidal ideation often use social media to express their feelings [12, 41], with eight out of ten people disclosing their suicidal thoughts and plans on social media [23].

[1]Code is available at https://github.com/midas-research/sismo-wsdm

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03…$15.00
https://doi.org/10.1145/3437963.3441805

The encouraging aspect is that progress in Machine Learning and Natural Language Processing (NLP) can complement social media analysis to identify risk markers in online user behavior [13] to aid suicide risk assessment [14]. However, language is subjective, and often, analyzing individual user posts may not be sufficient to assess a user's mental state and infer the associated suicide risk [24]. Promisingly, features such as a user's social network [36], or historic posts [34] can add context for analyzing a user's online behavior and have improved the predictive power of NLP models. The automatic risk assessment algorithms can now dramatically outperform prior, more traditional clinical prediction methods [12, 31].

The challenging aspect is that employing these methods in practice is not trivial. Several existing methods treat suicide risk assessment as a binary classification task [8, 32] that leads to the simplification of actual suicide risk levels. Such simplification can lead to artificial notions of risk [7], with no information on which users should be prioritized for clinical assessment. The lack of a finer-grained evaluation for the degree of risk poses a challenge: binary classifiers would flag a multitude of users when deployed, likely exceeding the limited capacity of a severely resource-limited mental health ecosystem [45, 49]. Recent advances in suicide risk assessment [21, 49] have shown the effectiveness of finer-grained assessment of at-risk users across increasing levels of suicide risk. We present a supporting example in Figure 1, where we show how such an assessment on a scale of increasing suicide risk of a user and attention could provide additional information to support the prioritization of high-risk users for clinical interventions. Despite these advancements, a limitation of existing approaches is they treat all the risk-levels equally, ignoring the inherent **ordinal** nature between risk-levels. That is, 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. In the event of not yielding the correct prediction, incorporating the ordinal relationship among suicide risk-levels can help predict a risk-level as close as possible to the actual risk.

Contributions: We reformulate suicide risk assessment as an ordinal regression problem, where *not all wrong classes are equally wrong* to prioritize at-risk users based on the increasing order of suicide risk levels **(Sec. 3.1)**. In this ordinal formulation, we map categorical risk levels to soft probability distributions to learn the natural inter-class relationships between suicide risk levels. To this end, we propose **SISMO**: **S**uicide **I**deation detection on **S**ocial **M**edia using an **O**rdinal formulation, an ordinal hierarchical attention model for suicide risk assessment. Through a series of quantitative **(Sec. 5)** and qualitative **(Sec. 5.3)** analyses on real *Reddit* data **(Sec. 4.1)**, we show that SISMO significantly outperforms state-of-the-art methods. Finally, we discuss the practical and ethical considerations pertaining to SISMO's usage as a preliminary tool for suicide risk assessment as part of a human-in-the-loop system **(Sec. 6)**.

At a minimum, we establish validity for formulating suicide ideation detection on social media as an ordinal regression task. We focus on the intersection of NLP and suicidal risk assessment by taking a step towards improving risk assessment in a non-intrusive manner. Our work *could* be used as a preliminary screening tool that optimistically forms a component in a larger infrastructure involving psychologists, health care providers, and social media enterprises.[2] In practice, SISMO could categorize potentially at-risk users based on C-SSRS for suicidality as part of a human-in-the-loop system to support decisions about potential intervention and the prioritization of clinical resources.

2 RELATED WORK

Neural Methods: Progress in Natural Language Processing has led to a rise in language-based suicide ideation detection on social media [30]. These early attempts analyze user features [33], online suicide notes [39], presence of psycho-linguistic lexicons such as LIWC [14] and textual features such as POS, tense, etc. for risk assessment [13]. However, these are language-specific methods that analyze individual posts in isolation and do not incorporate any additional user associated context. To gain a deeper insight into the user's mental state, studies have utilized user associated contexts such as a user's emotion spectrum, social graph methods [36] and the temporal context [42, 43, 45] of a user in the form of all historical posts. For instance, [8] used FastText embeddings of a user's posts and processed them sequentially using an LSTM to capture the temporal dependencies in the user's history. [46] used a stacked ensemble approach that utilizes the temporal context of a user along with their social graph. Despite showing performance improvements over non-contextual models, these models cannot support decisions regarding the prioritization of high-risk users for clinical interventions due to their binary classification scheme.

Fine-Grained Assessment: Recently, the focus of research has shifted to predicting suicide risk across varying degrees of severity [21, 34, 49]. Although the levels of risk severity lie on an ordinal scale, existing methods treat each risk level independently, implying that all incorrect predictions are equally wrong. However, in reality, it is desirable to predict a risk level that is as close as possible to the actual risk level in the case of an incorrect prediction. While ordinal regression has shown promise in automatic suicide risk assessment using the patient's non-textual clinical history [48], there are several limitations. First, clinical data contains large temporal gaps between clinical visits [12], and the samples are often homogeneous, which may not be representative of a larger population [13]. Further, the recording procedures may change because of the change of healthcare policies and the update of diagnosis codes [30], limiting historic clinical records usage. SISMO builds on these limitations and presents an ordinal formulation for risk assessment on social media.

3 METHODOLOGY

3.1 Problem Formulation

Our goal is to assess the suicide risk of a user $u_i \in U = \{u_1, u_2, \ldots, u_n\}$ by analyzing the posts $P_i = \{p_1^i, p_2^i, \ldots, p_{T_i}^i\}$ made by them on social media over time. Here, the posts P_i made by user u_i form a chronological sequence, based on their posting time, such that we denote p_t^i as the t^{th} post made by user u_i, with $p_{T_i}^i$ representing the most recent post. We formalize user assessment based on an adaptation of the Columbia-Suicide Severity Rating Scale (C-SSRS) [40] for social media [21]. Under the C-SSRS, a user may be classified in one of five categories of increasing risk: Support *(SU)* < Indicator

[2] Similar to screening deployed on Facebook. about.fb.com/suicide-prevention-and-ai

(IN) < Ideation *(ID)* < Behavior *(BR)* < Attempt *(AT)*. Given there exists a relative ordering between suicide risk levels, we formulate suicide risk assessment as an **Ordinal Regression** problem. Such an ordinal regression formulation essentially maps into a multi-class classification task where suicide risk levels are broken into multiple ranks, corresponding to the severity of the risk [19].

Formally, given a user u, having authored posts P over time, we aim to classify user u into one of five levels of increasing suicide risk: $y \in \{SU, IN, ID, BR, AT\}$. Our broader objective through this work is to design a data-driven model to assist the prioritization of at-risk users by clinicians, and relevant healthcare stakeholders through an assessment framework, such as one shown in Figure 1. We now present the components and optimization of SISMO.

3.2 Post Embedding

Each post made by a user could potentially be indicative of suicide risk, and provide context towards learning comprehensive representations of their mental health [13]. As shown in Figure 2, SISMO's first layer consists of a user post embedding layer that leverages word-level attention. Building on the success of transfer learning and pre-training of language models in NLP, we use Longformer [2], a pre-trained transformer language model to obtain post-level embeddings. Longformer's self attention mechanism emphasizes on tokens that can help discover risk-markers indicative of suicidality in a post. We tokenize each historical post, and add `[CLS]` token at the beginning of each post. We use the final hidden state corresponding to this `[CLS]` token as the aggregate representation of the post for further processing.[3] We encode each post p_t^i as $\mathbf{e}_t^i = \texttt{Longformer}(p_t^i)$.

3.3 User Context Modeling

The build up to suicide ideation can occur weeks, months or even years before its onset [38]. While each post could be indicative of suicidal risk, often analyzing the sequence of user posts can provide a deeper insight into their mental state and how it varies over time. A holistic representation of a user's mental state can highlight variations in potential risk markers over time [37]. We use a Bi-LSTM to capture temporal features across posts, that sequentially models post embeddings $\mathbf{e}^i$ for each post p^i authored by user u_i.

Sequential Context Modeling: The ability of Long-Short Term Memory networks (LSTMs) to capture long term dependencies [26] makes them suitable for modeling a user's posts chronologically on social media. For each sequential post p_t^i, a post encoding $\mathbf{e}_t^i$ is transformed into a contextualized representation $\mathbf{h}_t^i$ by gaining context from historic well as future posts i.e., concatenating the left-to-right and the right-to-left hidden state vectors of the LSTM.

$$\overrightarrow{\mathbf{h}^i}_t = \text{LSTM}(\mathbf{e}_t^i, \overrightarrow{\mathbf{h}^i}_{t-1}) \tag{1}$$

$$\overleftarrow{\mathbf{h}^i}_t = \text{LSTM}(\mathbf{e}_t^i, \overleftarrow{\mathbf{h}^i}_{t+1}) \tag{2}$$

$$\mathbf{h}_t^i = [\overrightarrow{\mathbf{h}^i}_t, \overleftarrow{\mathbf{h}^i}_t] \tag{3}$$

[3]We also experimented with the average of the output vectors over all tokens from Longformer's final layer, but empirically found that the `[CLS]` performed better.

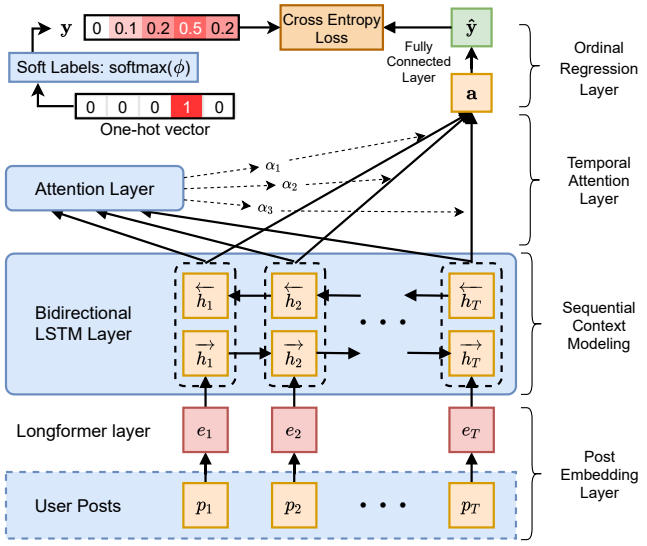


Figure 2: SISMO: Model Overview and Components

The BiLSTM transforms historical post encodings $[\mathbf{e}_1^i, \cdots, \mathbf{e}_{T_i}^i]$ into contextual representations $[\mathbf{h}_1^i, \cdots, \mathbf{h}_{T_i}^i] \in \mathbb{R}^{H \times T_i}$ where H is the dimension of each hidden state vector.

Temporal Attention: Often in a voluminous (possibly longitudinal) set of social media posts, only a few posts contain relevant signals.[4] As the degree of suicidality and presence of suicide association varies over each user post [22], we propose a temporal attention mechanism. This mechanism learns adaptive weights for contextual representations of each post ($\mathbf{h}_t^i$), highlighting posts with indicative markers for suicide risk and aggregates them as:

$$\mathbf{a}_i = \sum_{t=1}^{T_i} \alpha_t^i \mathbf{h}_t^i, \;\; \alpha_t^i = \frac{exp(\tilde{\alpha}_t^i)}{\sum_{t=1}^{T_i} exp(\tilde{\alpha}_t^i)} \tag{4}$$

$$\tilde{\alpha}_t^i = \mathbf{c}^T tanh(\mathbf{W}_x \mathbf{h}_t^i + \mathbf{b}_x) \tag{5}$$

where $\mathbf{W}_x \in \mathbb{R}^{T_i \times H}$, $\mathbf{b}_x \in \mathbb{R}^{T_i}$ and $\mathbf{c} \in \mathbb{R}^{T_i}$ are network parameters and $\mathbf{a}_i$ is the contextual representation of a user's historical posts.

3.4 Ordinal Regression

The final layer is a fully connected layer which takes the attention based representation $\mathbf{a}_i$ of user posts P_i made by user u_i as an input. This layer outputs classification confidence for all suicide risk-levels (λ) leveraging the ordinal ranking across risk groups, $\hat{\mathbf{y}}_i = \mathbf{W}_y \mathbf{a}_i + \mathbf{b}_y$ where, $\mathbf{W}_y \in \mathbb{R}^{\lambda \times H}$ and $\mathbf{b}_y$ are model parameters .

To preserve the natural ordering of severity of risk levels, we train SISMO by minimizing an ordinal regression loss [15]. It is similar to the cross-entropy loss used by suicide risk classification models that treat all the risk-levels equally. However, instead of using a one-hot vector representation of the ground truth, we use a soft encoded vector (probability distribution) representation along with the classification score $\hat{\mathbf{y}}$ to compute cross entropy loss.

[4]For example, [45], report that for a 'severe-risk' user, experts identified only two out of 1,326 *Reddit* postings displayed signals relevant to assess suicide risk.

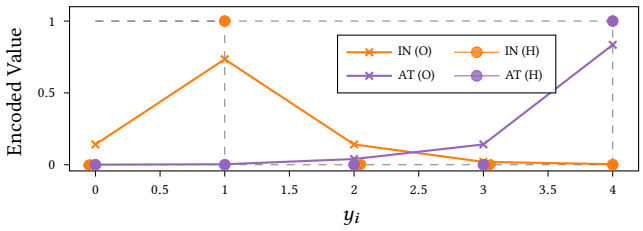


Figure 3: Probability distribution y of soft labels ($\alpha = 1.8$) (O) and one-hot vector (H) for different true risk-levels r_t. For $r_t = 4$, the probability y_i gradually decreases for corresponding risk-level r_i as we move away from the true risk-level, while using soft labels. In contrast, the one-hot vectors do not differentiate between the mispredicted classes.

Let $\mathcal{Y} = \{\text{SU} = 0, \text{IN} = 1, \text{ID} = 2, \text{BR} = 3, \text{AT} = 4\} = \{r_i\}_{i=0}^{4}$ represent the five ordinal risk-levels in the increasing order of severity. For a particular true risk-level $r_t \in \mathcal{Y}$, we compute soft labels as probability distributions $\mathbf{y} = [y_0, y_1, \cdots, y_4]$ of ground truth labels. The probability y_i of each risk-level r_i is:

$$y_i = \frac{e^{-\phi(r_t, r_i)}}{\sum_{k=1}^{\lambda} e^{-\phi(r_t, r_k)}} \; \forall r_i \in \mathcal{Y} \tag{6}$$

where $\phi(r_t, r_i)$ is a cost function that penalizes how far the true risk-level r_t is from a risk-level $r_i \in \mathcal{Y}$.

As the difference between a risk-level r_i and the true risk-level r_t increases, the probability y_i decreases for corresponding risk-level r_i as illustrated in Figure 3. After computing the probability distribution $\mathbf{y}$ for the ground truth label, we finally calculate the cross-entropy loss using classification score $\hat{\mathbf{y}}$ as:

$$\mathcal{L} = -\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{\lambda} \mathbf{y_{ij}} \log \hat{\mathbf{y}}_{\mathbf{ij}} \tag{7}$$

where n is the batch size and λ is the number of risk-levels (5).

The cost function ϕ is a pre-calculated inter-class cost penalizing the predictions far away from the true risk-level. We select ϕ such that predicting a user labeled "Actual Attempt (AT)" (highest risk category) by experts as Supportive (SU) should be strongly penalized as opposed to predicting Suicidal Behaviour (BR) which is closer to AT on the C-SSRS. We choose[5] $\phi(r_t, r_i) = \alpha|r_t - r_i|$, where α is a parameter that controls the magnitude of penalty for incorrectly predicting a risk-level.

4 EXPERIMENTAL SETUP

4.1 Data

We use the dataset released by [21] that consists of reddit posts of 500 users across 9 mental health and suicide related subreddits.[6] Although [21] started out with a dataset of 270,000 users, they filtered out users through various methods. Firstly, they applied a negation detection tool to filter trivially non-suicidal users. They further filtered user posts to reduce content overlap across different *subreddits*. Finally, they identified a cohort of 2181 *Redditors* as potential candidates for suicidal users from which 500 were chosen at random for annotation of suicide risk level across 5 categories of increasing risk. Annotation was performed by four practicing psychiatrists following guidelines as per the C-SSRS leading to an acceptable average pairwise agreement of 0.79 and group-wise agreement of 0.73. The detailed description of these classes along with examples has been provided in table 1. We observe that the number of tokens and 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 going upto a maximum of 292 posts. Similarly, the number of tokens in each post, is 73.4 ± 97.7. We summarize the risk levels (class labels) and their distribution with examples in Table 1.

Exploratory SAGE Analysis: In Table 1, we also explore the dataset to assess the language variation across the five levels of increasing suicide risk, using Sparse Additive Generative Model (SAGE) [17] analysis.[7] SAGE can be viewed as a combination of topic models and generalized additive models. SAGE utilizes the measure of a log-odds ratio to contrast word distributions between a corpus of interest against a baseline corpus, implying that the distinguishing words we obtain for a class are *relative* to all other classes. SAGE's additive nature allows us to see which words contribute most to each risk-level. The cluster of words for the Supportive level, consists of positive words such as *"perseverance"* and *"aspiration"*, as expected. As we move towards levels indicating higher suicide risk, such as Suicide Ideation, we notice clear signs of distress with the distinct usage of words like *"sleeplessness"* and *"tears"*. The users categorized in high risk levels, *Suicidal Behavior* and *Actual Attempt*, show explicitly suicidal tendencies with words such as *"blades"*, *"razor"* and *"handgun"*. As indicated by the most salient words in each risk category, there is a variation in the language of users across increasing levels of suicide risk.

Preprocessing and Data Split: We deidentified the dataset by performing named entity recognition and removing any identifiable information such as email addresses, URLs and names. Next, we follow standard procedures of converting the text to lowercase, removing punctuation and accents, striping whitespaces and removing stopwords. Following [21], we perform a stratified 80:20 split such that the train and test set consist of 400 and 100 users respectively. All experiments are performed with 5 fold hold-out cross-validation on the train set, with each cross-validation set consisting of 80 users. The final results are reported on the initially held-out test set. Multiple runs for each experiment are carried out with a different random initialization.

4.2 Evaluation Metrics

To better evaluate the models from the perspective of suicide risk assessment, we adopt the metrics proposed by [21]. They alter the formulation of False Negatives (FN) and False Positives (FP). FN is modified as the ratio of number of times predicted severity of suicide risk level (k^p) is less than the actual risk level (k^a) over the size of test data (N_T). FP is the ratio of number of times the predicted risk k^p is greater than the actual risk k^a. Precision and

[5] We tried other cost functions ϕ used in Computer Vision [15], but the improvements obtained were not statistically significant compared to $\phi(r_t, r_i) = \alpha|r_t - r_i|$.

[6] https://github.com/AmanuelF/Suicide-Risk-Assessment-using-Reddit

[7] github.com/jacobeisenstein/SAGE

Risk Level (Class-wise Distribution)	Definition	Example	SAGE Analysis
Supportive [SU] (20%)	Users who engage in discussion without using language indicative of being at-risk in the past or the present.	I *** tell you love **ur family... ... so ***eniably much... You care, *** ...that s**ws tr*m*ndo** str**gth.	perseverance 3.20 aspirations 3.08 paths 2.85
Suicidal Indicator [IN] (20%)	Users who use at-risk language but are not actively experiencing any general or acute symptoms.	I have had a *** history with anxiety... I get **** it's r**ll* difficult... The **rst step is ** always talk ** someone ab*** it.	engagement 1.12 impulse 1.12 foolish 0.88
Suicidal Ideation [ID] (34%)	Users who express thoughts of suicide or show pre-occupations with known risk factors.	I k*** Im depressed, have been... for m**t of ** life.... what's the **int of continu**g... it's just s**fering...	sleeplessness 1.68 tears 1.47 negativity 1.37
Suicidal Behavior [BR] (15%)	The confession of active or historical self-harm.	I pus**d all my frie**s and family away... ...I'm *** g**ng to drink... and drug myself t*** I pass out ... ** die al***.	blades 2.29 disowned 1.61 chronic 1.41
Actual Attempt [AT] (9%)	The confession of any deliberate action that could have resulted in intentional death, be it a completed attempt or not.	Ive tak** a bottle of pills before..., ...i try ** choke myself *v*r* few days... ...but **ver could fin**h i* ...	razor 2.49 isolate 1.51 handgun 1.31

Table 1: We define the categories under the adapted C-SSRS scheme of suicidal risk assessment and its class-wise distribution in the dataset. Alongside this, we provide an example of a user post associated with that suicide risk severity level. The posts have been paraphrased and obfuscated for user privacy. The top 3 most distinguishing words for each risk level are obtained using SAGE. A higher SAGE coefficient for a particular word is indicative of its saliency within the corpus of that risk level.

recall are reformulated as Graded Precision (GP) and Graded Recall (GR) to incorporate the ordering amongst risk levels.

$$FN = \frac{\sum_{i=1}^{N_T} I(k_i^a > k_i^p)}{N_T}, FP = \frac{\sum_{i=1}^{N_T} I(k_i^p > k_i^a)}{N_T}$$

where $\Delta(k_i^a, k_i^p)$ is the difference between actual k_i^a and predicted k_i^p suicide risk levels for user u_i.

4.3 Baselines

We compare SISMO with the following baselines:

Handcrafted Features: Approaches that concatenate the language based features used for suicide ideation detection such as Part of Speech counts, psychological LIWC features [13], and TF-IDF of all posts by a user into a single language vector.

- **SVM+RBF [1] :** Language vector for each user is fed to a Support Vector Machine with a Radial Basis Function kernel with the kernel parameter $\sigma = 0.24$ and the cost parameter $c = 5$.
- **SVM-L [1] :** A Support Vector Machine with a linear kernel ($c = 1.5$) takes the language vector for each user as input.
- **Random Forest [44] :** We applied Random Forest (500 trees, 9 predictor variables per tree node) to the language vectors.
- **MLP [1] :** Input vector fed to an MLP with two hidden layers with 64 neurons each.

Deep Learning Approaches.

- **Contextual CNN [21]:** Posts are encoded by concatenating GloVe word embeddings for each post. Bag of post embeddings are concatenated and fed to a contextual CNN.
- **Suicide Detection Model (SDM) [8]:** Fine-tuned FastText embeddings are used for encoding posts. Post embeddings are fed to an LSTM layer followed by an attention layer.
- **ContextBERT [34]:** The best performing model at the CL Psych 2019 shared task [49], ContextBERT uses BERT for encoding Reddit posts. The obtained BERT embeddings are then fed to a Gated Recurrent Unit.

4.4 Experimental Settings

We select hyperparameters based on the highest FScore obtained through cross-validation for all the models. We use grid search to explore: number of features in hidden state $H \in \{8, 64, 128, 256, 512\}$, number of LSTM layers $n \in \{1, 2, 5\}$, dropout $\delta \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, initial learning rate $lr \in \{0.01, 0.001, 0.0005, 0.0001\}$, and control parameter $\alpha \in \{0, 0.2, ..., 3.8\}$. The optimal hyperparameters were found to be $H = 512$, $n = 1$, $\delta = 0.3$, $lr = 0.01$ and $\alpha = 1.8$.

We fine-tune the base version of Longformer using Huggingface's *Transformers* library.[8] The language model and parameters are jointly trained. All the methods are implemented with PyTorch 1.5 and optimized using mini-batch AdamW with a batch size of 8. We handle the varying lengths of user histories by storing the sequence of post embeddings as a packed padded sequence. We train the model on an Nvidia Tesla K80 GPU for 50 epochs and apply early stopping with a patience of 5 epochs. We report the mean testing performance when SISMO performs best on the cross-validation set over five different runs, each with a different random initialization.

5 RESULTS

5.1 Performance Comparison

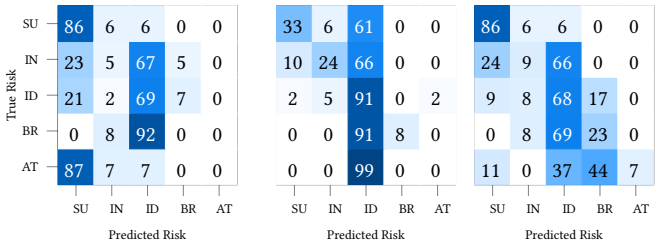
How does SISMO perform compared to the baselines?

We evaluate the model performance over five runs. Table 2 shows that SISMO significantly ($p < 0.05$) outperforms competitive baselines. We observe that deep learning approaches outperform the handcrafted feature-based models such as the SVM, Random Forest, and MLP. We believe this is because CNN/LSTM based models using embedding approaches can aggregate user posts more effectively to model the mental state of a user. Within deep learning models, we see that sequential models such as SDM and ContextBERT perform better than the contextual CNN. We postulate this to the ability of sequential models to learn better representations from the temporal

[8] https://huggingface.co/allenai/longformer-base-4096

Context Modeling	Language Features	Objective Function	Model	Graded Precision ↑	Graded Recall ↑	FScore ↑
Concatenated Bag-of-posts	Language-Based Features	Hinge Loss	SVM+RBF[1]	0.53	0.51	0.52
		Hinge Loss	SVM-L[1]	0.60	0.45	0.52
		Gini Impurity	Random Forest[1]	*0.68*	0.49	0.57
		Log Loss	MLP[1]	0.45	0.59	0.51
	Glove Embeddings	Cross Entropy	Contextual CNN[21]	**0.69**	0.52	0.59
Time Ordered	FastText Embeddings	Cross Entropy	SDM[8]	0.60	0.54	0.57
	BERT Embeddings	Cross Entropy	*ContextBERT[34]*	0.63	*0.57*	*0.60*
Sequential Hierarchical Attention	Longformer Embeddings	Ordinal Soft Label	**SISMO (Ours)**	0.66	**0.61***	**0.64***

Table 2: We report the median of results over 50 different runs. * indicates the result is statistically significant to Contextual CNN ($p < 0.05$) under Wilcoxon's signed rank test. Bold denotes best performance. *Italics* denotes second best. Results are highlighted with blue (better) and red (worse) with respect to the Contextual CNN. Best viewed in color.



(a) C-CNN **(b) C-BERT** **(c) SISMO**

Figure 4: Normalized Confusion matrix for the *median* of the test results. A row represents an instance of the actual risk-level, whereas a column represents an instance of the predicted risk-level. The values of the diagonal elements represent the percentage of correct predictions.

Model	Graded Precision ↑	Graded Recall ↑	FScore ↑
Contextual CNN	**0.83**	0.57	0.68
ContextBERT	0.79	0.66*	*0.71**
SISMO	0.81	**0.67***	**0.73***

Table 3: Results for 3+1 scheme. We report the median of results over 50 different runs. * indicates result is statistically significant compared to Contextual CNN ($p < 0.05$).

context in a user history in comparison to the Contextual CNN's bag-of-posts approach. ContextBERT shows a minor improvement over SDM, likely owing to the use of contextualized embeddings such as BERT over fixed word embeddings such as FastText. SISMO significantly outperforms ContextBERT, the current state-of-the-art model. We observe that SISMO performs better than ContextBERT, specifically, across the higher levels of risk, as shown in Figure 4. This improvement can be attributed to SISMO's ordinal formulation, which penalizes predictions based on how far the predicted risk is from the actual risk on the C-SSRS scale, as opposed to the baselines that treat each misclassification equally.

3+1 Label Classification: *How well does SISMO perform at a coarse grain segregation of high risk users compared to low (or no) risk users?*

Following [21], we observe results for an alternative scheme wherein we collapse the SU and IN classes into a common no-risk category. When we use this scheme, we observe an improvement in the precision of the predictive models from the 5 class scheme used earlier. We postulate this improvement to 1) the possibility of a reduction in the degree of freedom of the outcome variable (4 risk levels instead of 5), 2) as *supportive + indicator* and *ideation* classes are in majority, the models are more likely to predict the low-risk classes rather than *Behavior* and *Attempt*. We also observe from Table 3 that although the prediction for low-risk classes is similar across the models, SISMO can more correctly predict users at higher risk levels, which is reflected in the graded recall. .

5.2 Ablation Study

How does the model performance improve upon adding each component of SISMO to a naive BERT + Average Pooling model?

Model	Graded Precision ↑	Graded Recall ↑	FScore ↑
BERT+Average Pooling	0.56	0.59	0.57
BiLSTM	0.62	0.58	0.59
BiLSTM+Attn (SISM)	0.63	0.57	0.60
SISM+O (SISMO)	**0.66***	**0.59***	**0.64***

Table 4: Ablation results over SISMO's components. We report the median of results over 50 different runs. * indicates the result is statistically significant to BiLSTM ($p < 0.05$) under Wilcoxon's signed rank test.

We perform an ablation study to probe the effectiveness of each component of SISMO, as shown in Table 4. We add the sequential, temporal attention, and ordinal components to the base model (BERT+Average Pooling) that averages the BERT embeddings of all the posts made by a user. The sequential modeling of a user history shows an improvement over the base model (BERT) due to the ability of an LSTM to learn representations for temporal patterns from historical posts. We notice that the model performance improves on adding the temporal attention layer, possibly as it can identify certain posts in a user history that are more relevant for risk assessment (see Sec 5.3 for detailed examples). Once we utilize the relative ordering of risk levels, we observe a significant improvement in model performance.

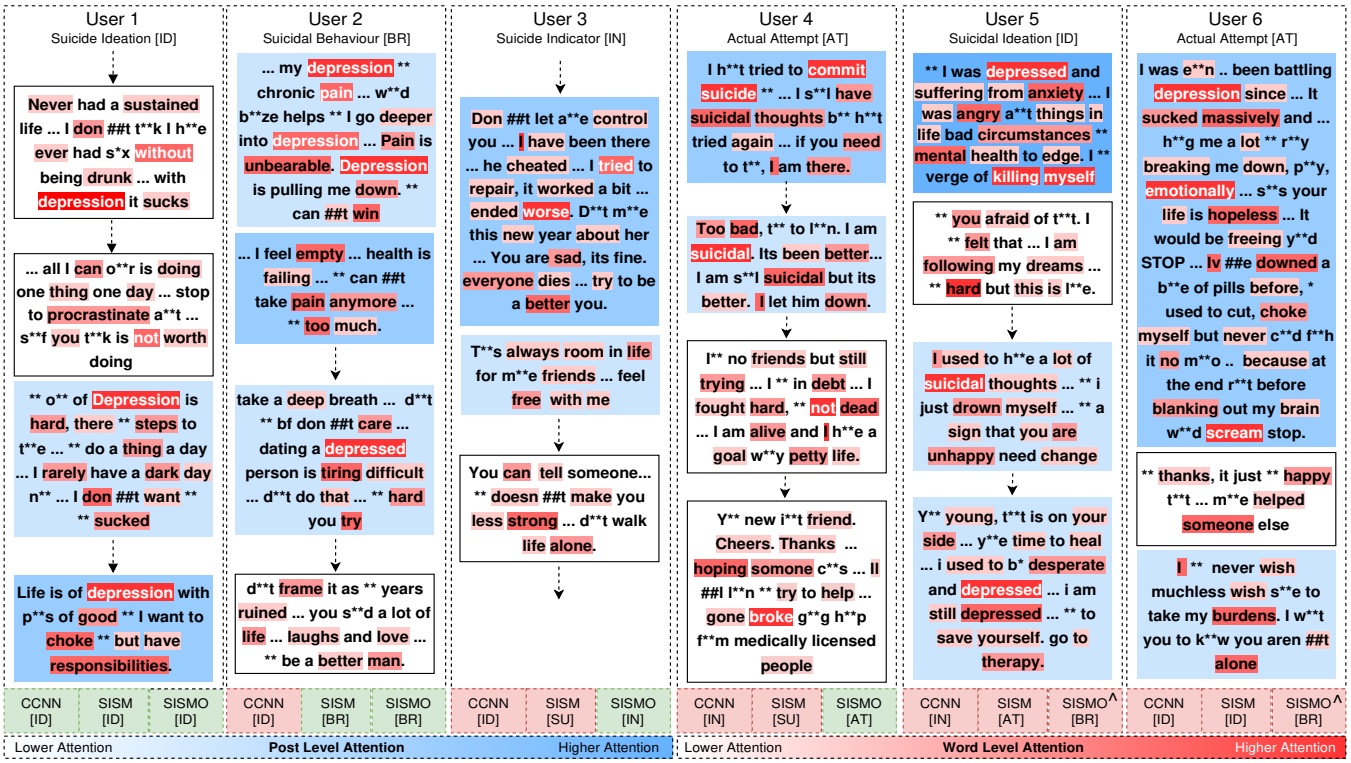


Figure 5: We show how SISMO can be interpreted for preliminary assessment of suicide-risk of users on social media. Significant posts (blue) are highlighted which, can be useful for review by a clinician and word-level attention (red) highlighting, foreground importance of information for severity of suicide risk assessment shown as per SISMO. The posts by the user are chronologically ordered, bottom to top (oldest at the bottom, latest on top). ↓ indicates posts not displayed for brevity. *Bottom:* We show predictions (green: correct, red: incorrect) made by models, ˆ denotes the closest prediction to the true risk.

5.3 Qualitative Analysis

How can we infer what the model is learning through actual examples?

We now present a qualitative analysis, as shown in Figure 5, that shows snippets from the posting histories of 6 users on *Reddit* along with the corresponding model predictions of Contextual CNN (CCNN), SISM, and SISMO. The token-level (red) and post-level (blue) attention assigned by SISMO are displayed, which can be indicative of the relevance of a word/post for the eventual prediction made by SISMO. We aim to derive insights into SISMO's predictive power in comparison to CCNN and SISM with the help of examples.

On analyzing the posts of user 1 and user 2, we see that user 1 consistently uses phrases indicative of suicide ideation like *"want to choke"*. In contrast, user 2 shows a build-up of sadness over time interspersed with a few supportive posts. Hence, for user 1, all three models can easily assess the associated risk-level correctly. Whereas for user 2, only SISM and SISMO identify the true level of risk. This observation signals the importance of sequentially modeling user history to extract temporal patterns in user posts.

User 3 belongs to the *indicator* class, which is a difficult risk-level to predict since the user tends to share their own experiences of suicidal thoughts to support others on the forum. Due to the presence of these opposing factors, CCNN and SISM, either underestimate or overestimate the user's risk. SISMO correctly predicts the risk level as *IN*, between the *SU* and *ID* risk categories likely owing to its ordinal formulation. Similarly, for user 4, we observe that SISMO learns attention weights that can identify specific posts displaying suicidal tendencies leading to a correct assessment of risk.

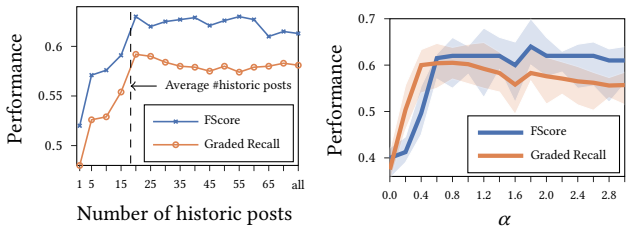
5.4 Error Analysis

When do modern deep learning models fail?

Although such NLP models show promise, they are susceptible to errors as well. For example, none of the models were able to assess the true risk for users 5 and 6 accurately. However, among the assessments made by the three models, the predictions of SISMO are closest to the true risk-level of the user. We attribute this to the ordinal regression component of SISMO that penalizes predictions based on the inherent increasing order of suicide risk. This property of SISMO resembles a human's judgment of risk [3] and is crucial in practical settings. SISMO classifies user 6 (AT) as BR which, is a more accurate assessment of the user's risk severity as compared to the low-risk predictions made by CCNN and SISM, which can impact the prioritization of user 6 in practice.

5.5 Parametric Analysis

How does varying user history and α affect SISMO's performance?



(a) Varying History **(b) Varying values of α**

Figure 6: FScore and Graded Recall of SISMO against (a) varying the number of posts in user history (vertical dashed line [- - -] represents average number of posts in a user history), and (b) varying values of control parameter α which controls the probability distribution of soft labels.

Results with Varying User History: We first discuss model performance given different amounts of posts in the user history. From Figure 6a, we can see that FScore and Graded Recall monotonically increases until we add 20 posts, which is the average number of historical posts for a user. When we further increase the number of posts in user history, the model performance does not significantly improve. We postulate this saturation in improvement to a *relatively* few number of users (9%) having more than 50 posts.

Results with Varying α: SISMO has one important parameter α, which controls the probability distribution of soft labels. We probe further to investigate the effect of this parameter by varying it from 0 to 3.0. In Figure 6b, we observe that when $\alpha = 0$ the model performance drastically degrades which, can be attributed to the uniform distribution of soft labels at $\alpha = 0$ as the model cannot differentiate between correct and wrong predictions. When α is too large ($a > 2$), the cost function ϕ heavily penalizes risk levels away from the true risk-level, forcing the soft labels to behave like one-hot labels. Hence, improving the performance over $\alpha = 0$. We observe that $\alpha = 1.8$ provides the best model performance by penalizing predictions based on how far the predicted risk is from the actual risk, as opposed to treating all misclassifications equally.

6 DISCUSSION

Acknowledging the sensitive nature of our work, we discuss the associated ethical implications, biases, and practical applicability.

Ethical Considerations: We work within the purview of acceptable privacy practices suggested by [10] and considerations discussed by [18] to avoid coercion and intrusive treatment. For the dataset [21] used in this research, the original *Reddit* data is publicly available. Although *Reddit* is intended for anonymous posting, we take further precautions by performing automatic de-identification of the dataset using named entity recognition [49]. Following [4], all example posts shown throughout this paper have been anonymized, obfuscated, and paraphrased for user privacy and to prevent misuse as per the *moderate disguise* scheme suggested by [6]. The annotation of user data has been kept separately from raw user data on protected servers linked only through anonymous IDs [4]. Our work focuses on developing a neural model for screening users and does not make any diagnostic claims related to suicide. We study *Reddit* posts in a purely observational capacity, and do not intervene with user experience. The assessments made by SISMO are sensitive and should be shared selectively and subject to IRB approval to avoid misuse like Samaritan's Radar [28].

Limitations: We acknowledge that the study of suicidal risk is subjective and the interpretation of the analysis presented may vary across individuals. 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 [5, 47]. We acknowledge that the studied data may be susceptible to demographic, expert annotator, and medium-specific biases [27] which can cause latent problems, especially when inferences are incorporated in real-life situations. The statistical patterns learned by SISMO may fail to generalize to different social networking websites and contexts, especially due to longer post length on Reddit. However, an interpretable model can help to track and rectify these different sources of statistical patterns and bias [29]. We also acknowledge that there exists a tradeoff between the inherent selection bias in the studied data and informed user consent [18].

Practical Applicability: Through SISMO, we suggest an interpretable neural architecture for **preliminary screening** of at-risk users on social media to aid the prioritization of clinical resources and users. A critical area following the task of screening is the design of an effective intervention. SISMO should form part of a human-centered mental health ecosystem involving psychiatrists, health-care providers, etc. Aside from clinical interventions, numerous stakeholders such as caregivers, close family members, and the at-risk individual can also be considered. There is also a need for cross-disciplinary collaboration and dialogue to ensure that there is no misuse and misinterpretation of the model's algorithmic outcomes. Following [11], we base our analysis on a well known clinical suicide risk scale, the C-SSRS, adhering to the standardization and well defined protocols for translating SISMO's predictions to clinical settings, and factoring in risk-dehumanization [9]. The presence of large online user histories, an aggregation of useful posts made by a user for downstream review by a clinician can improve SISMO's interpretability.

7 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 to prioritize at-risk users. In a broader view, ordinal regression solves the classification task for suicide risk-assessment where *not all wrong risk-levels are equally wrong*. We present SISMO, a dual attention hierarchical model for *Reddit* posts, to aid interpretability to assist clinicians in assessing voluminous user data. Through this work, we aim to form a component in a comprehensive human-in-the-loop infrastructure for preliminary suicide risk assessment of users on social media. Our future work includes ideating towards a more practical solution by collaborating with clinical psychologists, healthcare providers, and other stakeholders and taking additional measures for user privacy.

REFERENCES

[1] Payam Amini, Hasan Ahmadinia, Jalal Poorolajal, and Mohammad Moqaddasi Amiri. 2016. Evaluating the high risk groups for suicide: a comparison of logistic regression, support vector machine, decision tree and artificial neural network. *Iranian journal of public health* 45, 9 (2016), 1179.

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150* (2020).

[3] Ralf Bender and Ulrich Grouven. 1997. Ordinal logistic regression in medical research. *Journal of the Royal College of physicians of London* 31, 5 (1997), 546.

[4] Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. 94–102.

[5] Chloe Berryman, Christopher J. Ferguson, and Charles Negy. 2017. Social Media Use and Mental Health among Young Adults. *Psychiatric Quarterly* 89, 2 (Nov. 2017), 307–314. https://doi.org/10.1007/s11126-017-9535-6

[6] Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology* 4, 3 (2002), 217–231.

[7] Craig J Bryan and M David Rudd. 2006. Advances in the assessment of suicide risk. *Journal of clinical psychology* 62, 2 (2006), 185–200.

[8] Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention. In *Proceedings of EMNLP-IJCNLP 2019*.

[9] Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019. Who is the" Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on HCI* (2019).

[10] Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *In Proceedings of FAT*'. 79–88.

[11] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ* (2020).

[12] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights* 10 (2018), 1178222618792860.

[13] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *AAAI weblogs and social media 2013*.

[14] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings CHI 2016*. 2098–2110.

[15] Raul Diaz and Amit Marathe. 2019. Soft labels for ordinal regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[16] Christopher W Drapeau and John L McIntosh. 2020. USA suicide 2018: Official final data. (2020).

[17] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. (2011).

[18] Casey Fiesler and Nicholas Proferes. 2018. "Participant" perceptions of Twitter research ethics. *Social Media+ Society* 4, 1 (2018), 2056305118763366.

[19] Eibe Frank and Mark Hall. 2001. A Simple Approach to Ordinal Classification. In *Machine Learning: ECML 2001*, Luc De Raedt and Peter Flach (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 145–156.

[20] Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. 2017. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychological bulletin* (2017).

[21] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*. 514–525.

[22] Jeffrey J Glenn, Alicia L Nobles, Laura E Barnes, and Bethany A Teachman. 2019. Can Text Messages Identify Suicide Risk in Real Time? A Within-Subjects Pilot Examination of Temporally Sensitive Markers of Suicide Risk. *Clinical Psychological Science* (2019), 2167702620906146.

[23] Robert N Golden, Carla Weiland, and Fred Peterson. 2009. *The truth about illness and disease*. Infobase Publishing.

[24] Judith Rich Harris. 2010. *No two alike: Human nature and human individuality*. WW Norton & Company.

[25] Holly Hedegaard, Sally C Curtin, and Margaret Warner. 2020. Increase in suicide mortality in the United States, 1999–2018. (2020).

[26] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780.

[27] Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of ACL 2016 (Volume 2: Short Papers)*. 591–598.

[28] Honor Hsin, John Torous, and Laura Roberts. 2016. An Adjuvant Role for Mobile Health in Psychiatry. *JAMA Psychiatry* 73, 2 (Feb. 2016), 103.

[29] Nicholas C Jacobson, Ashley Walton, Alexander J Millner, Garth Coombs III, and Alexandra M Rodman. 2020. Ethical dilemmas posed by mobile health and machine learning in psychiatry research. *Bulletin of the WHO* (2020).

[30] Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2019. Suicidal ideation detection: A review of machine learning methods and applications. *arXiv preprint arXiv:1910.12611* (2019).

[31] Kathryn P Linthicum, Katherine Musacchio Schafer, and Jessica D Ribeiro. 2019. Machine learning in suicide science: Applications and ethics. *Behavioral sciences & the law* 37, 3 (2019), 214–222.

[32] David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of eRisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 340–357.

[33] Naoki Masuda, Issei Kurahashi, and Hiroko Onari. 2013. Suicide Ideation of Individuals in Online Social Networks. *PloS one* 8 (04 2013), e62262.

[34] Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019. Suicide Risk Assessment with Multi-level Dual-Context Language and BERT. In *Proceedings of CLPsych 2019*. ACL 2019, 39–44.

[35] Catherine M McHugh, Amy Corderoy, Christopher James Ryan, Ian B Hickie, and Matthew Michael Large. 2019. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. *BJPsych open* 5, 2 (2019).

[36] Rohan Mishra, Pradyumna Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, IIIT MIDAS, and Rajiv Ratn Shah. 2019. SNAP-BATNET: Cascading Author Profiling and Social Network Graphs for Suicide Ideation Detection on Social Media. *NAACL HLT 2019* (2019), 147.

[37] John L Oliffe, John S Ogrodniczuk, Joan L Bottorff, Joy L Johnson, and Kristy Hoyak. 2012. "You feel like you can't live anymore": Suicide from the perspectives of Canadian men who experience depression. *Social science & medicine* (2012).

[38] James Overholser. 2003. Predisposing factors in suicide attempts: life stressors. In *Evaluating and treating adolescent suicide attempters*. Elsevier, 41–52.

[39] John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide Note Classification Using Natural Language Processing: A Content Analysis. *Biomedical informatics insights* 2010 (2010).

[40] Kelly Posner, Gregory K. Brown, Barbara Stanley, David A. Brent, Kseniya V. Yershova, Maria A. Oquendo, Glenn W. Currier, Glenn A. Melvin, Laurence Greenhill, Sa Shen, and J. John Mann. 2011. The Columbia–Suicide Severity Rating Scale: Initial Validity and Internal Consistency Findings From Three Multisite Studies With Adolescents and Adults. *AJP* (2011).

[41] Jo Robinson, Georgina Cox, Eleanor Bailey, Sarah Hetrick, Maria Rodrigues, Steve Fisher, and Helen Herrman. 2016. Social media and suicide prevention: a systematic review. *Early intervention in psychiatry* 10, 2 (2016), 103–121.

[42] Ramit Sawhney, Harshit Joshi, Lucie Flek, and Rajiv Ratn Shah. 2021. PHASE: Learning Emotional Phase-aware Representations for Suicide Ideation Detection on Social Media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

[43] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7685–7697. https://doi.org/10.18653/v1/2020.emnlp-main.619

[44] Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. 2018. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*. 91–98.

[45] Han-Chin Shing, Philip Resnik, and Douglas W Oard. 2020. A Prioritization Model for Suicidality Risk Assessment. In *Proceedings of ACL 2020*. 8124–8137.

[46] Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. 2019. # suicidal-A Multipronged Approach to Identify and Explore Suicidal Ideation in Twitter. In *Proceedings of CIKM 2019*.

[47] Hajime Sueki. 2015. The association of suicide-related Twitter use with suicidal behaviour: A cross-sectional study of young internet users in Japan. *Journal of Affective Disorders* 170 (Jan. 2015), 155–160. https://doi.org/10.1016/j.jad.2014.08.047

[48] Truyen Tran, Dinh Phung, Wei Luo, Richard Harvey, Michael Berk, and Svetha Venkatesh. 2013. An integrated framework for suicide risk prediction. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1410–1418.

[49] Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of CLPsych 2019*. ACL 2019.