PHASE: Learning Emotional Phase-aware Representations for Suicide Ideation Detection on Social Media

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

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

Lucie Flek
Mainz University of Applied Sciences
lucie.flek@hs-mainz.de

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

Abstract

Recent psychological studies indicate that individuals exhibiting suicidal ideation increasingly turn to social media rather than mental health practitioners. Contextualizing the build-up of such ideation is critical for the identification of users at risk. In this work, we focus on identifying suicidal intent in tweets by augmenting linguistic models with emotional phases modeled from users' historical context. We propose PHASE, a time-and phase-aware framework that adaptively learns features from a user's historical emotional spectrum on Twitter for preliminary screening of suicidal risk. Building on clinical studies, PHASE learns phase-like progressions in users' historical Plutchik-wheel-based emotions to contextualize suicidal intent. While outperforming state-of-the-art methods, we show the utility of temporal and phase-based emotional contextual cues for suicide ideation detection. We further discuss practical and ethical considerations.[1]

1 Introduction

Every 10.9 minutes, a person dies of suicide (Drapeau and McIntosh, 2020). Suicide ranks as the second leading cause of death for 14-35 year-olds (Hedegaard et al., 2020) in US. Extending appropriate clinical and psychological care to suicidal people relies on identifying those at risk. Unfortunately, 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 (McHugh et al., 2019; Franklin et al., 2017). In contrast, people exhibiting suicidal ideation often use social media to express their feelings (Coppersmith et al., 2014, 2016, 2018; Robinson et al., 2016; Reger et al., 2020), with eight out of ten people disclosing their suicidal thoughts and plans on social media (Golden et al., 2009).

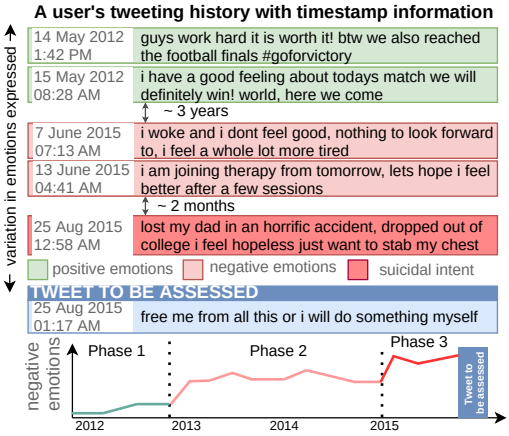

Figure 1: We study a user's tweeting history and emotional progression. Note that while the user's most recent tweet (blue) shows a subtle indication of suicidal intent, it is not sufficient to ascertain suicide risk. Grouping the build-up of negative emotions (red) in the user's historical tweets into phase-like emotional progressions, by utilizing the elapsed time between tweets, can contextualize the user's state and provide a more accurate and interpretable risk assessment. *All examples in this paper have been anonymized and paraphrased as per a moderate disguise scheme (Bruckman, 2002) to protect user privacy (Chancellor et al., 2019b).*

Natural Language Processing (NLP) presents an encouraging prospect to complement social science to identify risk markers in user behavior (De Choudhury et al., 2013, 2016) to aid suicide risk assessment (Shing et al., 2018, 2020). However, suicide ideation is complex, and often, individual posts may not be sufficient to assess a user's suicide risk, even for humans (Sisask et al., 2008; O'dea et al., 2015). Figure 1 illustrates how features such as historical posts (Matero et al., 2019) can add context for analyzing a user's online behavior over time (Van Heeringen and Marušic, 2003) to better

* Equal contribution

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

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2415–2428
April 19 - 23, 2021.

ascertain suicide risk. Despite the success of user-centric contextual models (Flek, 2020) for suicide ideation detection, they have two major limitations.

First, recurrent neural networks, particularly LSTMs, that are natural methods to learn patterns from a sequence of a user's historical tweets (Cao et al., 2019; Zeng et al., 2019, 2020), assume uniform time gaps between successive tweets. However, tweets can be posted at irregular time intervals (Lei et al., 2018), and varying time gaps can influence the assessment of a user's suicidality progression (Chen et al., 2018), as shown in Figure 1.

Second, these methods implicitly assume that a user's mental and emotional state progression is smooth in time, with an ever-increasing tendency. However, in reality, studies show that emotional (Larsen et al., 2015), and suicidality progression can vary significantly (Bryan and Rudd, 2016; Bryan, 2020), and show fluctuating *phase* like patterns (Kiosses et al., 2014; Palmier-Claus et al., 2012). Analyzing such phase-wise emotion progressions and build-up, as illustrated in Figure 1, can be instrumental in contextualizing suicidal risk, and aiding clinical psychologists through increased interpretability in human-in-the-loop systems.[2]

Building on these limitations, and motivated by psychological studies (Neacsiu et al., 2018; Domínguez and Fernández, 2018) of emotional state progression, we propose **PHASE: PH**ase-**A**ware **S**uicidality identification **E**motion progression model. With PHASE, we present the first neural framework to identify suicide ideation on social media (§**3.1**) that explicitly models the *inherent phase-aware progressions* in users' emotional spectrums in a contextual time-aware manner.

We present the following key contributions:

(i) First, building on the success of large scale pretraining in NLP, we utilize Plutchik Transformer, a transformer to learn linguistic and Plutchik-based (Plutchik, 1980) emotional cues from tweets (§**3.2**).

(ii) We propose Time-Sensitive Emotion LSTM (TSE-LSTM) to learn the historical emotional progression of a user's mental states from their learned emotional spectrum in a time-aware manner (§**3.3**).

(iii) Based on psychological studies, we propose a novel method to learn users' emotional phase progressions by leveraging the amount of historical emotional context used to update the TSE-LSTM's cell state. PHASE identifies the onset of new emotional phases and learns a temporal *phase-aware emotional user representation* (§**3.4**) that is then used to identify suicide ideation in their recent tweets (§**3.5**), increasing the system transparency.

(iv) Through a series of experiments (§**4.2**), we show that PHASE significantly ($p < 0.005$) outperforms competitive methods, which do not take users' emotional phases into account (§**5**).

(v) We analyze the contributions of PHASE's individual components to suicide ideation detection (§**5.2,** §**5.3,** §**5.4**), assess its transparency and limitations through qualitative analysis (§**5.5**), and conclude by discussing the ethical implications and practical applicability of this study (§**6**).

2 Related Work

Traditional Methods: Researchers have devised various psychoclinical methods to assess suicidal risk (Pestian et al., 2016), such as the Suicide Probability Scale (Bagge and Osman, 1998), Suicide Ideation Questionnaire (wa Fu et al., 2007), Suicidal Affect-Behavior-Cognition Scale (Harris et al., 2015). While these methods are professional and effective, they require participants to answer questionnaires (Venek et al., 2017) or engage in interviews (Scherer et al., 2013), hence not reaching people who cannot access these resources or have a low motivation to seek professional help (Zachrisson et al., 2006; Essau, 2005). Harris and Goh (2016) show that such assessments can negatively impact people showing depressive symptoms.

NLP Methods: Recently, social media has shown promise in providing insights into users' mental states (Paul and Dredze, 2011). Jashinsky et al. (2014) reported that Twitter is a viable tool for real-time monitoring (Braithwaite et al., 2016) of suicide risk. Early efforts in utilizing social media leverage user features such as their age, gender,and social network connectivity (Masuda et al., 2013) and online suicide notes (Pestian et al., 2010). Since then, the focus has been on using psycholinguistic lexicons such as LIWC and textual features such as n-grams, POS tags, etc. for classification (De Choudhury et al., 2016; Sawhney et al., 2018b). Shared tasks such as CLPsych (Zirikly et al., 2019) and CLEF eRISK (Losada et al., 2020) have seen a rise in neural networks such as CNNs (Yates et al., 2017; Du et al., 2018; Naderi et al., 2019; Gaur et al., 2019) and LSTMs (Ji et al., 2018; Tadesse et al., 2020) to predict suicide risk. While these methods capture post semantics in isolation, no user context is leveraged, hindering insight into

[2]Similar to the post-screening on Facebook (Card, 2018).

the user's mental state to improve predictive power (Venek et al., 2017; Flek, 2020). User context includes the user's emotions (Ren et al., 2016; Guntuku et al., 2017), social networks (Mishra et al., 2019) and historical posts (Mathur et al., 2020).

Contextual Methods: The best performing model, the DualContextBERT (Matero et al., 2019), at CLPsych 2019 for suicidal estimation exemplifies the utility of temporal context. The DualContextBERT models post embeddings sequentially via an RNN. Such RNN-based approaches assume that users' historical posts are equally spaced in time, hindering their ability to learn their relative importance in a time-aware manner. Recently, time-aware modeling of well defined stages in numerical time series data shows promising results in clinical tasks like patient subtyping (Baytas et al., 2017) and disease progression (Gao et al., 2020). However, the time-sensitive phase extraction of user-generated posts on social media, and phase-aware modeling of textual data is underexplored and complex, as it involves noisy, unstructured and ambiguous inputs across irregular time intervals.

3 PHASE: Components and Learning

3.1 Notations and Problem Formulation

We formulate suicidal intent detection as a binary classification task to predict suicidal intent y_i for a tweet t_i, where, $y_i \in \{$suicidal intent present, suicidal intent absent$\}$. We denote the tweet to be assessed for the presence of suicidal intent as $t_i \in T = \{t_1, t_2, \cdots, t_N\}$, authored by a user $u_j \in U = \{u_1, u_2, \cdots, u_M\}$, posted at time τ^i_{curr}. Each tweet t_i is associated with history $H^j_i = [(h^i_1, \tau^i_1), (h^i_2, \tau^i_2), \cdots, (h^i_L, \tau^i_L)]$ where h^i_k is a historic tweet authored by user u_j posted at time τ^i_k with $\tau^i_1 < \tau^i_2 < \cdots < \tau^i_L < \tau^i_{curr}$.

As shown in Figure 2, PHASE first obtains a user's emotion spectrum from their historical tweets and the tweet to be assessed using a fine-tuned BERT model, Plutchik Transformer. We feed the historical tweet representations to our proposed Time-Sensitive Emotion LSTM to learn the temporal progression of a user's emotions. We then identify *phases* in a user's emotions from their learned historical emotional progression, and extract temporal features for a user from these phases using Phase-Adaptive convolutions. Finally, PHASE jointly learns the semantics of user tweets and their historical emotional context in a temporal phase-aware manner for suicide ideation detection in a tweet.

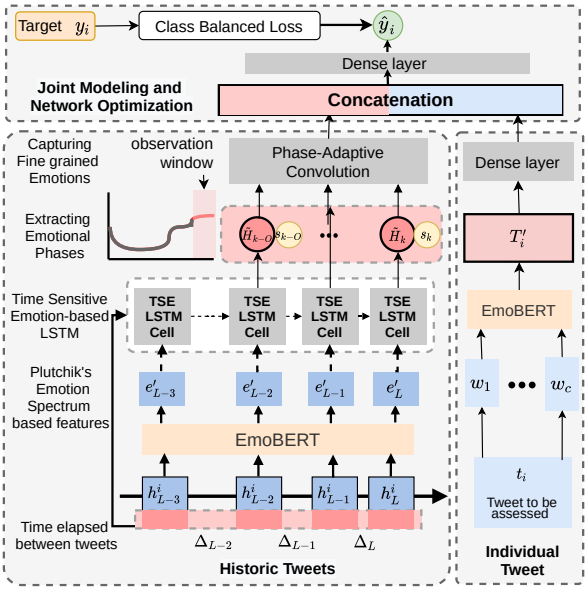

Figure 2: An illustration of PHASE's architecture.

3.2 Plutchik Transformer: Encoding Tweets

Studies show that emotions expressed in suicidal tweets are correlated with suicidal behavior (Sueki, 2015; Spates et al., 2018; Zhang et al., 2017). As a building block, we utilize Plutchik's wheel of emotions (Plutchik, 1980) to capture the emotions expressed by a user in their tweets. Plutchik's wheel outlines eight primary emotions arranged as four pairs of opposing dualities: Joy - Sadness, Surprise - Anticipation, Anger - Fear, and Trust - Disgust. We utilize Plutchik Transformer (Sawhney et al., 2020), a BERT model fine-tuned on Emonet (Abdul-Mageed and Ungar, 2017), a dataset of 790,059 tweets labeled across 8 primary emotions as per Plutchik's wheel of emotions. Owing to the success of pre-training language models in NLP, Plutchik Transformer jointly learns textual and emotion features for representation learning of user tweets for subsequent suicidal intent detection. We extract a 768-dimension encoding from the `[CLS]`[3] token of the penultimate transformer layer, which is densely connected with an 8-dimensional output layer representative of each primary emotion.

[3]Empirically, the `[CLS]` token performed better than taking the average of the output vectors over all tokens.

Tweet to be assessed: We encode each tweet to be assessed t_i as:

$$\mathbf{T_i'} = \texttt{PlutchikTransformer}(t_i) \quad (1)$$

where $\mathbf{T'_i} \in \mathbb{R}^{768}$ is linearly transformed using a dense layer to $\mathbf{T_i} \in \mathbb{R}^d$ with dimension d.

Historical Tweet Encoding: A holistic representation of users' emotional states can be indicative of variations in risk markers over time (Aragón et al., 2019; Tarrier et al., 2007; Links et al., 2008). To this end, we utilize Plutchik Transformer to encode each historical tweet h_k^i to an emotion representation ($\mathbf{e_k^i} \in \mathbb{R}^{768}$) defined as:

$$\mathbf{e_k^i} = \texttt{PlutchikTransformer}(h_k^i) \quad (2)$$

3.3 Temporal Modeling of Historical Tweets

Building on these natural irregularities in posting times of historical tweets (Wojcik and Hughes, 2019), we propose the use of ON-LSTM (Shen et al., 2018) to encode the sequence of a user's historical tweet emotion representations $\mathbf{e_k^i}$ to capture the variation in their mental and emotional states over time, forming a Time-Sensitive Emotion LSTM (TSE-LSTM). In our TSE-LSTM, we introduce a time-sensitive long-term gate $\tilde{\mathbf{f}}_\mathbf{k}$, which contains older historic emotional context. Additionally, we propose a short-term gate $\tilde{\mathbf{i}}_\mathbf{k}$ that encodes recent historic tweets, as shown in Figure 3. We then feed the time-lapsed Δ_k from the previous tweet and the historical emotional representation $\mathbf{e_k^i}$ of each tweet h_k^i to a TSE-LSTM cell. This design aids TSE-LSTM to learn two probability distributions $\mathbf{p}_{\tilde{\mathbf{f}}_\mathbf{k}}$ and $\mathbf{p}_{\tilde{\mathbf{i}}_\mathbf{k}}$ corresponding to the long-term and short-term gates, respectively. Psychological studies show that a user's recent emotions can be more indicative of their current mental state (Fawcett et al., 1990; Homan et al., 2014). To this end, we set the update frequency of the short-term gate higher than the long-term gate to increase the influence of their more recent emotional context. To impose this natural ordering of frequency updates, we apply cumulative sum ($cumsum$) operation to the probability distributions $\mathbf{p}_{\tilde{\mathbf{f}}_\mathbf{k}}$ and $\mathbf{p}_{\tilde{\mathbf{i}}_\mathbf{k}}$:

$$\mathbf{p}_{\tilde{\mathbf{f}}} = \sigma(\mathbf{W}_{\tilde{\mathbf{f}}}(\mathbf{e_k^i} \oplus \Delta_k) + \mathbf{U}_{\tilde{\mathbf{f}}}(\tilde{\mathbf{H}}^\mathbf{i}_{\mathbf{k-1}} \oplus \Delta_k) + \mathbf{b}_{\tilde{\mathbf{f}}}) \quad (3)$$

$$\mathbf{p}_{\tilde{\mathbf{i}}} = \sigma(\mathbf{W}_{\tilde{\mathbf{i}}}(\mathbf{e_k^i} \oplus \Delta_k) + \mathbf{U}_{\tilde{\mathbf{i}}}(\tilde{\mathbf{H}}_{\mathbf{k-1}} \oplus \Delta_k) + \mathbf{b}_{\tilde{\mathbf{i}}}) \quad (4)$$

$$\tilde{\mathbf{f}}_\mathbf{k} = \overrightarrow{cumsum}(\mathbf{p}_{\tilde{\mathbf{f}}}), \tilde{\mathbf{i}}_\mathbf{k} = \overleftarrow{cumsum}(\mathbf{p}_{\tilde{\mathbf{i}}}) \quad (5)$$

where σ represents softmax, $\oplus$ denotes concatenation and $\tilde{\mathbf{H}}^\mathbf{i}_{\mathbf{k-1}}$ is the previous hidden state.

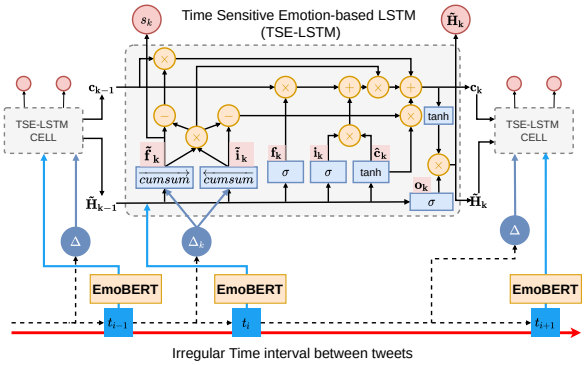

Figure 3: Detailed structure of the TSE-LSTM cell. Figure is adapted from (Gao et al., 2020).

The arrow above $cumsum$ indicates its direction. $\mathbf{W}_{\tilde{\mathbf{f}}}, \mathbf{W}_{\tilde{\mathbf{i}}}, \mathbf{U}_{\tilde{\mathbf{f}}}, \mathbf{U}_{\tilde{\mathbf{i}}}, \mathbf{b}_{\tilde{\mathbf{f}}}$ and $\mathbf{b}_{\tilde{\mathbf{i}}}$ are learnable parameters. Following $cumsum$'s properties, the values in the long-term gate $\tilde{\mathbf{f}}_\mathbf{k}$ are monotonically increasing from 0 to 1, and those in the short-term gate $\tilde{\mathbf{i}}_\mathbf{k}$ are monotonically decreasing from 1 to 0.

For each historic tweet h_k^i, the long-term gate $\tilde{\mathbf{f}}_\mathbf{k}$ controls the historic emotional context to be discarded, and the short-term gate $\tilde{\mathbf{i}}_\mathbf{k}$ controls the importance of recent historic emotions. To obtain complete contextual information of overlapping context in $\tilde{\mathbf{f}}_\mathbf{k}$ and $\tilde{\mathbf{i}}_\mathbf{k}$, we introduce a historic overlap vector $\mathbf{w_k}$ that uses the standard forget and input gates, $\mathbf{f_k}$ and $\mathbf{i_k}$, respectively. We define the new update function for TSE-LSTM's cell state $\mathbf{c_k}$ as:

$$\hat{\mathbf{c}}_\mathbf{k} = tanh(\mathbf{W_c}\mathbf{e}^\mathbf{i}_{\mathbf{k-1}} + \mathbf{U_c}\tilde{\mathbf{H}}^\mathbf{i}_{\mathbf{k-1}} + \mathbf{b_c}) \quad (6)$$

$$\mathbf{w_k} = \tilde{\mathbf{f}}_\mathbf{k} \odot \tilde{\mathbf{i}}_\mathbf{k} \quad (7)$$

$$\mathbf{c_k} = \mathbf{w_k} \odot (\mathbf{f_k} \odot \mathbf{c_{k-1}} + \mathbf{i_k} \odot \tilde{\mathbf{c}}_\mathbf{k}) + (\tilde{\mathbf{f}}_\mathbf{k} - \mathbf{w_k}) \odot \mathbf{c_{k-1}} + (\tilde{\mathbf{i}}_\mathbf{k} - \mathbf{w_k}) \odot \hat{\mathbf{c}}_\mathbf{k} \quad (8)$$

$$\tilde{\mathbf{H}}^\mathbf{i}_\mathbf{k} = \mathbf{o_k} \odot tanh(\mathbf{c_k}) \quad (9)$$

where computation for the intermediate cell state $\hat{\mathbf{c}}_\mathbf{k}$, output gate $\mathbf{o_k}$, the hidden state $\tilde{\mathbf{H}}^\mathbf{i}_\mathbf{k}$ are the same as in the standard LSTM and $\mathbf{W_c}$, $\mathbf{U_c}$, $\mathbf{b_c}$ are network parameters. The hidden state $\tilde{\mathbf{H}}^\mathbf{i}_\mathbf{k}$ represents the learned emotional context of the user.

3.4 Learning Emotional Phase Progression

We now describe how we use the emotional context learned by the TSE-LSTM to capture emotional phase progression patterns for a user over time. We then describe PHASE's Phase Adaptive Convolutions (PACs) that capture user features closely related to the user's current state through convolutions over these learned emotional phases. The PACs thus extract a phase-aware emotional user representation for suicide ideation detection.

Emotion Phase Variation: We leverage the historical emotional context $\tilde{\mathbf{H}}^{\mathbf{i}}_{\mathbf{k}}$ from the TSE-LSTM to extract temporal variations in a user's emotional state for a macroscopic view of the progression of *emotional phases*. Building on the work of Gao et al. (2020), we capture the onset of a new emotional phase by observing the proportion of historic context discarded to update the cell state $\mathbf{c_k}$. When almost no historical emotional context is used to update the cell state c_k, we say that a new emotional phase of the user has begun. Formally, we use a phase split point (s_k) that represents the time, before which all the emotional historic context is discounted ($\mathbf{p}_{\tilde{\mathbf{f}}_\mathbf{k}}$), as $s_k = argmax(\mathbf{p}_{\tilde{\mathbf{f}}})$, as shown in Figure 4 (Gao et al., 2020). Intuitively, a large value of s_k means little historic context is used to update the state cell c_k, indicating the onset of a new emotional phase; whereas, a smaller value of s_k suggests a long-term dependency of the emotions expressed in the tweet (h^i_k) on historic emotions. Since $argmax$ is non-differentiable, we estimate the phase split point (s_k) as:

$$s_k \approx \sum_{i=1}^{N_h} i \times \mathbf{p}_{\tilde{\mathbf{f}}}(i) = N_h \left(1 - \frac{1}{N_h} \sum_{i=1}^{N_h} \tilde{\mathbf{f}}_\mathbf{k}(i)\right) + 1 \quad (10)$$

where N_h is the dimension of $\tilde{\mathbf{H}}^{\mathbf{i}}_{\mathbf{k}}$, $\tilde{\mathbf{f}}_\mathbf{k}(i)$ and $\mathbf{p}_{\tilde{\mathbf{f}}}(i)$ are i^{th} values in the long-term gate $\tilde{\mathbf{f}}_\mathbf{k}$, and $\mathbf{p}_{\tilde{\mathbf{f}}}$.

We then compute the elapsed time between two consecutive phases by measuring the difference between the proportion of historic context discarded at each timestep. For each emotional phase of a user within an observation window of length L^w, we define this phase variation time $\mathbf{\Delta s}$ as:

$$\mathbf{\Delta s} = \sigma(\overrightarrow{cumsum}(s_{k-L^w}, \cdots, s_k)) \quad (11)$$

Phase Adaptive Convolution (PAC): We now extract features from the emotional phase build-up leading towards the tweet to be assessed. The PAC extracts features from the learned phase-wise progression of a user's temporal emotional context in the most recent emotional phase, as shown in Figure 4. We feed the concatenated historical hidden states $\tilde{\mathbf{H}}^{\mathbf{i}}_{\mathbf{k-L^w:k}} = [\tilde{\mathbf{H}}_{\mathbf{k-L^w}}, \cdots, \tilde{\mathbf{H}}^{\mathbf{i}}_{\mathbf{k}}]$, in the observation window L^w, as an input to a weighted temporal convolution. Naturally, emotions corresponding to more recent phases of a user are more indicative of their current mental state, and should be more influential (Larsen et al., 2009). Hence, we weigh the importance of the learned historical emotional context through the phase variation time

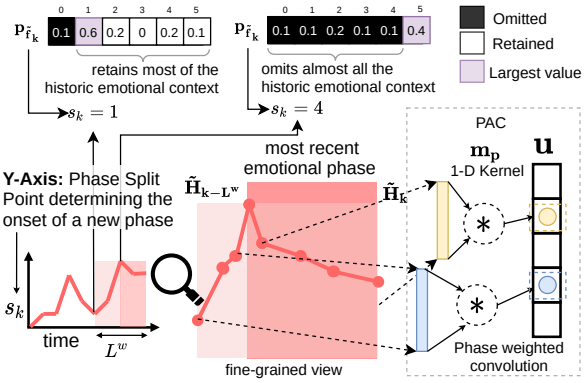

Figure 4: The Phase-Adaptive Convolution takes the historical hidden state $\tilde{\mathbf{H}}^{\mathbf{i}}_{\mathbf{k}}$ and the phase split-point s_k within the observation window length L^w as an input and learns a user's representation through their temporal patterns ($\mathbf{u}$), using phase-weighted convolutions.

Δs (Gao et al., 2020). We perform a convolution with a p^{th} 1-dimensional learnable kernel (m^j_p) for each j^{th} hidden state in the observation window as:

$$\mathbf{u^p} = \mathbf{m_p} * \tilde{\mathbf{H}}_{\mathbf{k-L^w:k}} = \sum_{j=1}^{N_h} \mathbf{m^j_p} * (\tilde{\mathbf{H}}^{\mathbf{j}}_{\mathbf{k-L^w:k}} \odot \mathbf{\Delta s}) \quad (12)$$

where $*$ is convolution operation, $\mathbf{u_p}$ is output of p^{th} kernel of size L^w. We concatenate all extracted features as $\mathbf{u} = [\mathbf{u^1}, \cdots, \mathbf{u^{N_h}}] \in \mathbb{R}^{N_h}$ to obtain a user's phase-aware emotion representation.

3.5 PHASE Joint Network Optimization

Finally, we concatenate encoded representations of the tweet to be assessed $\mathbf{T_i}$ and the historic emotional context $\mathbf{u}$, followed by softmax (σ) over a dense layer with a Rectified Linear Unit ($ReLU$).

$$\hat{\mathbf{y}}_\mathbf{i} = \sigma(ReLU(\mathbf{W_y}(\mathbf{T_i} \oplus \mathbf{u}) + \mathbf{b_y})) \quad (13)$$

where $\hat{\mathbf{y}}_\mathbf{i}$ is the final suicide risk assessment and $\{\mathbf{W_y}, \mathbf{b_y}\}$ are learnable network parameters.

Tweets with SI present form a very small proportion of the data (Ji et al., 2019). To address this problem of class imbalance *(the imbalance is much greater in the real world)*, we train PHASE using Class-Balanced Focal Loss (Lin et al., 2017; Cui et al., 2019). This loss function re-weights loss inversely with the effective number of samples per class, thereby yielding a class-balanced loss $\mathcal{L}$ as:

$$\mathcal{L} = \text{CB}_{focal}(\hat{\mathbf{y}}_\mathbf{i}, y_i; \beta, \gamma) \quad (14)$$

where CB_{focal} is class-balanced focal loss, $\hat{\mathbf{y}}_\mathbf{i}$ is the predicted label and y_i is the label of the tweet to be assessed. β and γ are hyperparameters.

4 Experimental Setup

4.1 Data and Preprocessing

We build on an existing Twitter data curated by Mishra et al. (2019). The data includes 34,306 tweets authored by 32,558 unique users. These tweets were identified based on a lexicon of 143 suicidal phrases (e.g., "wanting to die", "last day"). Two students of Psychology annotated the data under the supervision of a professional clinical psychologist, achieving a Cohen's Kappa score of 0.72, under the below guidelines (Sawhney et al., 2018b):
Suicidal Intent (SI) Present: Tweets where suicide ideation or previous attempts are discussed in a somber and non-flippant tone.
Suicidal Intent (SI) Absent: Tweets with no evidence for risk of suicide, e.g., song lyrics, condolence message, awareness, news.
The resulting dataset contains 3984 suicidal tweets. The Twitter timeline was collected for each user, spanning over ten years from 2009 to 2019. The number of historical tweets (748 ± 789) and the time difference between consecutive tweets (2 ± 24 days) are indicative of large variations across users. 4070 users were found to have no historical tweets.

We perform a stratified 70:10:20 split, such that the train, validation, and test sets consist of 24014, 3431, and 6861 tweets, respectively, and ensure that there is no overlap between users in these sets.

4.2 Baselines and Training Setup

We evaluate PHASE on macro F1 and recall (SI present) with both tweet- and user-level methods.

Tweet-level Non-contextual Baselines
RF + TF (Sawhney et al., 2018b): Extracts features including statistical, LIWC features, n-grams (up to 4), and POS counts from the tweet to be assessed and feeds them to a Random Forest (RF) classifier.
C-LSTM (Sawhney et al., 2018a): A deep neural network having a CNN followed by an LSTM to extract short and long range features in a tweet.

User-level Contextual Baselines
C-CNN (Gaur et al., 2019): A model that is fed GloVe encoded tweets as a concatenated bag of tweets, non-sequentially to a contextual CNN (Shin et al., 2018) with max pooling (Shing et al., 2018).
Suicide Detection Model (SDM) (Cao et al., 2019): Historical tweets encoded using fine-tuned FastText embeddings are fed to a regular LSTM followed by a tweet-level attention mechanism.
DualContextBert (Matero et al., 2019): Best performing model at CLPsych 2019. BERT embeddings of each historical tweet are sequentially fed to a regular RNN followed by tweet-level attention.
Exponential Decay (Sinha et al., 2019): A deep neural network that models encodes each historical tweet using Glove embeddings followed by a BiLSTM with attention. The historical embeddings are then aggregated using an exponential decay.

Setup: We set hyperparameters for all models based on the validation macro F1 score. We use grid search to explore: $N_h \in \{128, 256, 512\}$, $\delta \in \{0.0, 0.1, \cdots, 0.5\}$, $\beta \in \{0.99, 0.999, 0.9999\}$ and $\gamma \in \{1.0, 1.5, 2.0\}$, initial learning rate $I_{lr} \in \{0.01, 0.001, 0.0001\}$, $L^w \in \{1, 2, \cdots, 16\}$. We found the optimal hyperparameters as: N_h=512, δ=0.5, β =0.9999, γ=2, I_{lr}=0.0001, L^w=5. We implement all methods with PyTorch 1.6, and optimize PHASE using AdamW with a batch size of 128 for 30 epochs in 167 mins on a Tesla K80 GPU. We use the cosine scheduler (Gotmare et al., 2018) with a warmup step of 5.

5 Results and Analysis

5.1 Performance Comparison

Model	Macro F1↑	Recall$_s$↑	Acc↑
RF+TF	0.513	0.536	0.548
C-LSTM	0.588	0.597	0.602
C-CNN	0.729	0.587	0.803
SDM	0.743	0.755	0.819
DualContextBert	0.767	0.786	0.823
Exponential Decay	0.737	0.759	0.828*
PHASE	**0.805***	**0.812***	**0.856***

Table 1: Median of results over 5 different runs. * indicates improvement over DualContextBert is significant ($p < 0.005$) under Wilcoxon's Signed Rank test.

We observe from Table 1 that PHASE significantly ($p < 0.005$) outperforms all baselines. We note that contextual models outperform the non-contextual RF+TF and C-LSTM, as they learn a holistic representation of a user's mental state. Amongst contextual models, we note that models that factor in the temporal sequence of historical tweets outperform the non temporal C-CNN, that models tweets as a bag-of-tweets. Thereby validating the utility of temporal context for suicide ideation detection. PHASE significantly outperforms state-of-the-art contextual models. We postulate this to PHASE's ability to capture irregularities in tweeting patterns and learning emotional phase

PHASE (Ablative) Components	F1↑	Recall_s↑
Current Tweet only (C)	0.731	0.597
C + History (HST) + LSTM	0.780	0.794
C + HST + TSE-LSTM	0.796*	0.788
PHASE:C+HST+TSE-LSTM+PAC	**0.805***	**0.812***

Table 2: * shows significant improvements compared to C + HST + LSTM ($p < 0.005$). We use Plutchik Transformer as the tweet encoder for comparison.

progressions, unlike DualContextBERT and SDM, that ignore both the time- and phase-sensitive and emotional aspects of historical context. PHASE outperforms Exponential Decay, as PHASE adaptively learns progressions of emotional phases, rather than assuming a user's behavior to follow a specific trajectory that might not generalize well across users. This observation is in line with psychological research (Joiner Jr, 2002; Giletta et al., 2015) that shows the progressive build-up to suicidality varies across individuals, that PHASE is able to capture better than competitive models.

5.2 PHASE Components Ablation Study

We perform an ablation study to probe the effectiveness of each component of PHASE, as shown in Table 2, starting from the base (Current) model that does not use historical tweets. On modeling the temporal dependencies in historical tweets with a standard LSTM along with the current tweet, we note drastic improvements, revalidating the prominence of user-level context to infer the suicidality of a user. We then observe that on factoring in time-sensitivity through the TSE-LSTM, there is a significant ($p < 0.005$) improvement in the macro F1 score, but there is no gain in Recall. We believe even though the model gains additional user context by factoring in the time irregularities between tweets, the model does not improve drastically, as it still assumes a continuous smooth progression of the user's emotions in time. This assumption hinders the model's ability to capture the macroscopic context acquired by analyzing the phase like progressions of a user's emotional states (Homan et al., 2014). On adding the PAC that learns phase-aware user representations by extracting emotional progression patterns from their historical emotional context (TSE-LSTM), we observe significant ($p < 0.005$) improvements. We attribute this improvement to the PAC as it adaptively learns and captures a user's emotional phase-wise build-up towards their most recent tweet to be assessed, to correctly contextualize suicidal intent, validating the utility of phase-aware modeling.

5.3 Probing Plutchik Transformer: Encoder Analysis

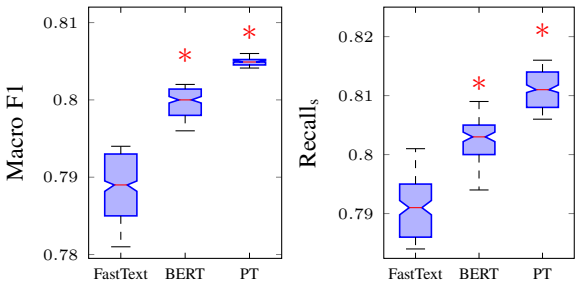

Figure 5: Performance with different tweet encoders over 10 different runs. PT: Plutchik Transformer. All improvements ($*$) are significant ($p < 0.005$) under Wilcoxon's Signed Rank test.

We now analyze PHASE's performance using different encoders to learn representations for tweets. Overall, we observe that transformers outperform previously used static word embeddings (FastText). Additionally, we observe that Plutchik Transformer, based on Plutchik's wheel of emotions, significantly improves PHASE's performance over the pre-trained BERT used by Matero et al. (2019). This observation revalidates the importance of specific emotional context, as opposed to the more general language features learned by BERT alone.

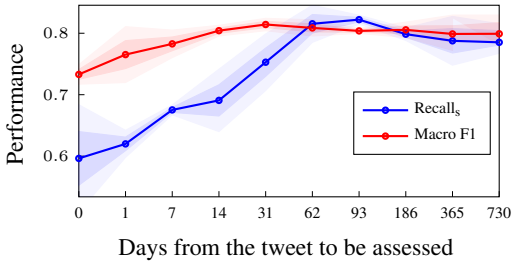

Figure 6: Influence of historical context (upto d days)

5.4 How much Historical Context is useful?

We explore PHASE's performance variation with the amount of historical lookback in terms of number of days in Figure 6. We observe that PHASE's performance improves as we factor in more historical tweets, going back up to a few months, likely as PHASE gains more context of users' emotional progressions. As we further increase historical lookback beyond several (> 3) months, we observe that PHASE's performance saturates. This observation is in line with psychological studies (Selby et al.,

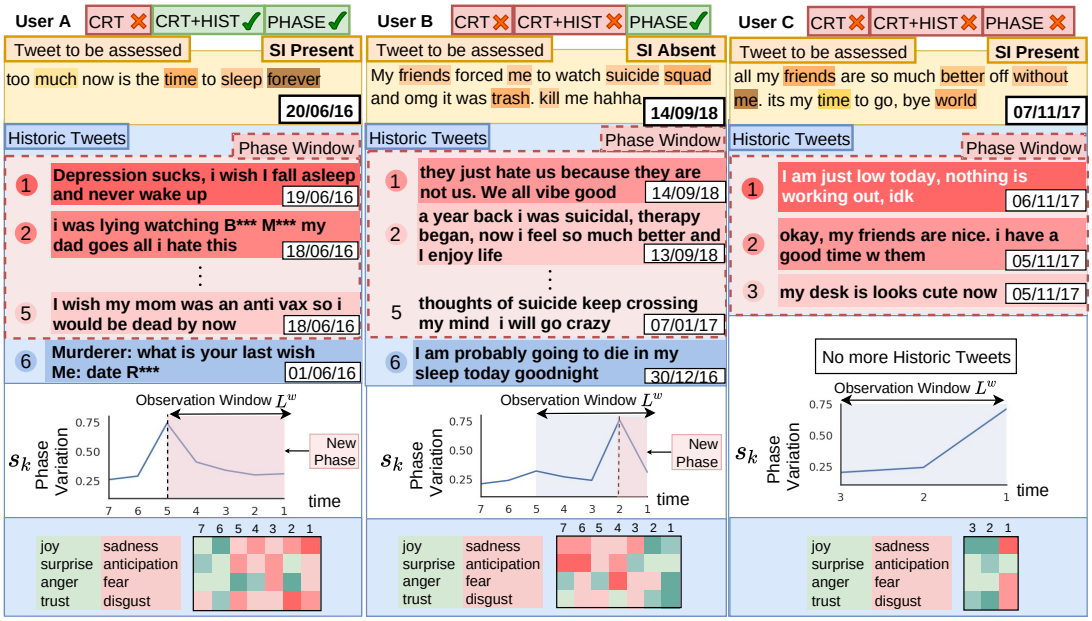

Figure 7: We study three users with their tweet to be assessed, historic tweets (chronologically ordered), and timestamps, showing how PHASE can aid human moderators and clinical psychologists with explainable predictions. We visualize self-attention (averaged over all 12 Plutchik Transformer heads) per token, where darker intensity denotes higher attention. The graphs show the phase split value s_k for each user over time. We also show emotional phase progression for further interpretability, where a peak represents the onset of a new phase. Further, we show detailed phase variation by visualizing the Plutchik-based emotion (learned weights) duality for historical tweets.

2013; Kaplow et al., 2014; Glenn et al., 2020), that highlight the diminishing importance of a user's emotions over longer time periods in assessing their current mental state and associated suicide risk.

5.5 PHASE Analysis and Interpretation

We now analyze PHASE's preliminary assessment in Figure 7 to elucidate on PHASE's interpretability to aid subsequent human-in-the-loop risk assessment. First, for User A, we see no apparent signs of suicidal intent in their tweet to be assessed, and if analyzed in isolation, is not sufficient to ascertain risk. However, User A's historical tweets add context to models (e.g. PHASE) that leverage temporal emotional cues to identify suicidal intent correctly. Next, we analyze a complex case, User B, where we observe *phase*-like progressions in their emotions over time. Although User B historically did show negative emotions, recently, User B shows more positive behavior, akin to the onset of a new emotional phase characterized by joy and trust. PHASE's design enables it to learn User B's emotional progression adaptively and correctly predicts User B's tweet to be analyzed as having no suicidal intent, unlike other models that incorrectly assess this as a tweet having suicidal intent. Lastly, we also present the complicated case of User C, where all models fail to explicate the future challenges in online data-driven suicide ideation. Specifically, we find that all models are unable to accurately ascertain suicidal intent where there is little historical context consisting of fluctuating emotions, highlighting the challenges associated with new or alternate accounts of users, amongst other complexities (Shea, 1999; O'Connor and Portzky, 2018).

6 Ethical and Practical Considerations

Emphasizing the sensitive nature of this work, we acknowledge the trade-off between privacy and effectiveness (Eskisabel-Azpiazu et al., 2017), and utilize publicly available Twitter data in a purely observational (Norval and Henderson, 2017; Broer, 2020), and non-intrusive manner. We separate user data from all other data on protected servers linked only through anonymous IDs, and we perform automatic de-identification of the dataset using named entity recognition (Benton et al., 2017a,b). All examples shown in this work have been paraphrased to protect user privacy (Fiesler and Proferes, 2018; Chancellor et al., 2019a,b). We ensure that this analysis is shared selectively and subject to IRB approval (Zimmer, 2009, 2010) to avoid misuse such as Samaritan's Radar (Hsin et al., 2016). We acknowledge that suicidality is subjective (Keilp

et al., 2012) and that the interpretation of this analysis may vary across individuals (Puschman, 2017). We further acknowledge that suicide risk exists on a diverse spectrum (Bryan and Rudd, 2006), rather than at a binary level, and that the studied data may be susceptible to demographic, annotator, and medium-specific biases (Hovy and Spruit, 2016). Finally, our work does not make any diagnostic claims related to suicide. PHASE should form part of a distributed human-in-the-loop (de Andrade et al., 2018) system for finer interpretation of risk.

7 Conclusion

Motivated by the rising exhibition of suicide ideation on social media, we present PHASE. Building on psychological studies analyzing the emotional spectrum and mental health of users, PHASE adaptively learns emotional phase-aware user representations through historical tweeting activity for suicidal ideation detection. We propose multiple modeling innovations in PHASE components: contextualized historical emotion representations (Plutchik Transformer), time-sensitive emotion LSTM (TSE-LSTM), and a phase-adaptive convolution (PAC). We demonstrate that modeling user phases explicitly increases the predictive power in assessing suicidality in tweets. In a qualitative analysis, we show how PHASE can aid social media moderators and clinical psychologists in subsequent assessment by displaying its predictions together with the learned emotional phases. Through PHASE, we hope to form a future component in a larger human-in-the-loop infrastructure for suicide prevention.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. 2018. Ethics and artificial intelligence: Suicide prevention on facebook. *Philosophy & Technology*, 31(4):669–684.

Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes-y Gómez. 2019. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1481–1486, Minneapolis, Minnesota. Association for Computational Linguistics.

Courtney Bagge and Augustine Osman. 1998. The suicide probability scale: Norms and factor structure. *Psychological Reports*, 83(2):637–638.

Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 65–74.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017a. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017b. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. 2016. Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR Mental Health*, 3(2):e21.

Tineke Broer. 2020. Technology for our future? exploring the duty to report and processes of subjectification relating to digitalized suicide prevention. *Information*, 11(3):170.

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):217–231.

Craig J. Bryan. 2020. Chapter 4 - the temporal dynamics of the wish to live and the wish to die among suicidal individuals. In Andrew C. Page and Werner G.K. Stritzke, editors, *Alternatives to Suicide*, pages 71 – 88. Academic Press.

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

Craig J Bryan and M David Rudd. 2016. The importance of temporal dynamics in the transition from suicidal thought to behavior. *Clinical Psychology: Science and Practice*, 23(1):21–25.

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 the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1718–1728, Hong Kong, China. Association for Computational Linguistics.

Catherine Card. 2018. How Facebook AI helps suicide prevention. *Facebook Newsroom*.

Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. 2019a. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32.

Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019b. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 79–88, New York, NY, USA. Association for Computing Machinery.

Xuetong Chen, Martin D. Sykora, Thomas W. Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1653–1660, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10:117822261879286.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 106–117, San Diego, CA, USA. Association for Computational Linguistics.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

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 of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Elena Domínguez, García and Pablo Fernández, Berrocal. 2018. The association between emotional intelligence and suicidal behavior: a systematic review. *Frontiers in psychology*, 9:2380.

Christopher W Drapeau and John L McIntosh. 2020. Usa suicide 2018: Official final data.

Jingcheng Du, Yaoyun Zhang, Jianhong Luo, Yuxi Jia, Qiang Wei, Cui Tao, and Hua Xu. 2018. Extracting psychiatric stressors for suicide from social media using deep learning. *BMC medical informatics and decision making*, 18(Suppl 2):43–43.

Amaia Eskisabel-Azpiazu, Rebeca Cerezo-Menéndez, and Daniel Gayo-Avello. 2017. An ethical inquiry into youth suicide prevention using social media mining. *Internet Research Ethics for the Social Age*, 227.

Cecilia A. Essau. 2005. Frequency and patterns of mental health services utilization among adolescents with anxiety and depressive disorders. *Depression and Anxiety*, 22(3):130–137.

Jan Fawcett, William A Scheftner, Louis Fogg, David C Clark, Michael A Young, Don Hedeker, and Robert Gibbons. 1990. Time-related predictors of suicide in major affective disorder. *The American journal of psychiatry*.

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366.

Lucie Flek. 2020. Returning the n to nlp: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838.

Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, 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*, 143(2):187.

King wa Fu, Ka Y. Liu, and Paul S. F. Yip. 2007. Predictive validity of the chinese version of the adult suicidal ideation questionnaire: Psychometric properties and its short version. *Psychological Assessment*, 19(4):422–429.

Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference 2020*, pages 530–540.

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*, pages 514–525.

Matteo Giletta, Mitchell J Prinstein, John RZ Abela, Brandon E Gibb, Andrea L Barrocas, and Benjamin L Hankin. 2015. Trajectories of suicide ideation and nonsuicidal self-injury among adolescents in mainland china: Peer predictors, joint development, and risk for suicide attempts. *Journal of consulting and clinical psychology*, 83(2):265.

Jeffrey J Glenn, Alicia L Nobles, Laura E Barnes, and Bethany A Teachman. 2020. Can text messages identify suicide risk in real time? a within-subjects pilot examination of temporally sensitive markers of suicide risk. *Clinical Psychological Science*, 8(4):704–722.

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

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. In *International Conference on Learning Representations*.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Keith M. Harris and Melissa Ting-Ting Goh. 2016. Is suicide assessment harmful to participants? findings from a randomized controlled trial. *International Journal of Mental Health Nursing*, 26(2):181–190.

Keith M. Harris, Jia-Jia Syu, Owen D. Lello, Y. L. Eileen Chew, Christopher H. Willcox, and Roger H. M. Ho. 2015. The ABC's of suicide risk assessment: Applying a tripartite approach to individual evaluations. *PLOS ONE*, 10(6):e0127442.

Holly Hedegaard, Sally C Curtin, and Margaret Warner. 2020. Increase in suicide mortality in the united states, 1999–2018.

Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. 2014. Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 107–117.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Honor Hsin, John Torous, and Laura Roberts. 2016. An adjuvant role for mobile health in psychiatry. *JAMA Psychiatry*, 73(2):103.

Jared Jashinsky, Scott H. Burton, Carl L. Hanson, Josh West, Christophe Giraud-Carrier, Michael D. Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through twitter in the US. *Crisis*, 35(1):51–59.

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*.

Shaoxiong Ji, Celina Ping Yu, Sai fu Fung, Shirui Pan, and Guodong Long. 2018. Supervised learning for suicidal ideation detection in online user content. *Complexity*, 2018:1–10.

Thomas E Joiner Jr. 2002. The trajectory of suicidal behavior over time. *Suicide and life-threatening Behavior*, 32(1):33–41.

Julie B Kaplow, Polly Y Gipson, Adam G Horwitz, Bianca N Burch, and Cheryl A King. 2014. Emotional suppression mediates the relation between adverse life events and adolescent suicide: Implications for prevention. *Prevention science*, 15(2):177–185.

John G. Keilp, Michael F. Grunebaum, Marianne Gorlyn, Simone LeBlanc, Ainsley K. Burke, Hanga Galfalvy, Maria A. Oquendo, and J. John Mann. 2012. Suicidal ideation and the subjective aspects of depression. *Journal of Affective Disorders*, 140(1):75–81.

Dimitris N Kiosses, Katalin Szanto, and George S Alexopoulos. 2014. Suicide in older adults: the role of emotions and cognition. *Current psychiatry reports*, 16(11):495.

Mark E Larsen, Tjeerd W Boonstra, Philip J Batterham, Bridianne O'Dea, Cecile Paris, and Helen Christensen. 2015. We feel: mapping emotion on twitter. *IEEE journal of biomedical and health informatics*, 19(4):1246–1252.

Randy J Larsen, Adam A Augustine, and Zvjezdana Prizmic. 2009. A process approach to emotion and personality: Using time as a facet of data. *Cognition and Emotion*, 23(7):1407–1426.

Kai Lei, Ying Liu, Shangru Zhong, Yongbin Liu, Kuai Xu, Ying Shen, and Min Yang. 2018. Understanding user behavior in sina weibo online social network: a community approach. *IEEE Access*, 6:13302–13316.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Paul S Links, Rahel Eynan, Marnin J Heisel, and Rosane Nisenbaum. 2008. Elements of affective instability associated with suicidal behaviour in patients with borderline personality disorder. *The Canadian Journal of Psychiatry*, 53(2):112–116.

David E Losada, Fabio Crestani, and Javier Parapar. 2020. erisk 2020: Self-harm and depression challenges. In *European Conference on Information Retrieval*, pages 557–563, Lisbon, Portugal. Springer.

Naoki Masuda, Issei Kurahashi, and Hiroko Onari. 2013. Suicide ideation of individuals in online social networks. *PloS one*, 8:e62262.

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 the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44.

Puneet Mathur, Ramit Sawhney, Shivang Chopra, Maitree Leekha, and Rajiv Ratn Shah. 2020. Utilizing temporal psycholinguistic cues for suicidal intent estimation. In *Lecture Notes in Computer Science*, pages 265–271. Springer International Publishing.

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).

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019. 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*, pages 147–156, Minneapolis, Minnesota. Association for Computational Linguistics.

Nona Naderi, Julien Gobeill, Douglas Teodoro, Emilie Pasche, and Patrick Ruch. 2019. A baseline approach for early detection of signs of anorexia and self-harm in reddit posts. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*, CONFERENCE. 9-12 September 2019.

Andrada D Neacsiu, Caitlin M Fang, Marcus Rodriguez, and M Zachary Rosenthal. 2018. Suicidal behavior and problems with emotion regulation. *Suicide and Life-Threatening Behavior*, 48(1):52–74.

Chris Norval and Tristan Henderson. 2017. Contextual consent: Ethical mining of social media for health research. *CoRR*, abs/1701.07765.

Bridianne O'dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

Rory C O'Connor and Gwendolyn Portzky. 2018. Looking to the future: a synthesis of new developments and challenges in suicide research and prevention. *Frontiers in Psychology*, 9:2139.

J. E. Palmier-Claus, P. J. Taylor, F. Varese, and D. Pratt. 2012. Does unstable mood increase risk of suicide?: theory, research and practice. *Journal of Affective Disorders*, 143(1-3):5–15.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*.

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(3):19–28.

John P. Pestian, Michael Sorter, Brian Connolly, Kevin Bretonnel Cohen, Cheryl McCullumsmith, Jeffry T. Gee, Louis-Philippe Morency, Stefan Scherer, and Lesley Rohlfs and. 2016. A machine learning approach to identifying the thought markers of suicidal subjects: A prospective multicenter trial. *Suicide and Life-Threatening Behavior*, 47(1):112–121.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Cornelius Puschman. 2017. Bad judgment, bad ethics? *Internet Research Ethics for the Social Age*, page 95.

Mark A. Reger, Ian H. Stanley, and Thomas E. Joiner. 2020. Suicide Mortality and Coronavirus Disease 2019—A Perfect Storm? *JAMA Psychiatry*.

Fuji Ren, Xin Kang, and Changqin Quan. 2016. Examining accumulated emotional traits in suicide blogs with an emotion topic model. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1384–1396.

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):103–121.

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)*, pages 7685–7697, Online. Association for Computational Linguistics.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018a. 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*, pages 167–175, Brussels, Belgium. Association for Computational Linguistics.

Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. 2018b. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98, Melbourne, Australia. Association for Computational Linguistics.

S. Scherer, J. Pestian, and L. Morency. 2013. Investigating the speech characteristics of suicidal adolescents. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 709–713.

Edward A Selby, Shirley Yen, and Anthony Spirito. 2013. Time varying prediction of thoughts of death and suicidal ideation in adolescents: weekly ratings over 6-month follow-up. *Journal of Clinical Child & Adolescent Psychology*, 42(4):481–495.

Shawn Christopher Shea. 1999. *The practical art of suicide assessment: A guide for mental health professionals and substance abuse counselors.* John Wiley & Sons Inc.

Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. 2018. Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*.

Joongbo Shin, Yanghoon Kim, Seunghyun Yoon, and Kyomin Jung. 2018. Contextual-cnn: A novel architecture capturing unified meaning for sentence classification. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 491–494. IEEE.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Han-Chin Shing, Philip Resnik, and Douglas W Oard. 2020. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137.

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 the 28th ACM International Conference on Information and Knowledge Management*, pages 941–950.

Merike Sisask, Airi Värnik, Kairi Kolves, Kenn Konstabel, and Danuta Wasserman. 2008. Subjective psychological well-being (who-5) in assessment of the severity of suicide attempt. *Nordic Journal of Psychiatry*, 62(6):431–435.

Kamesha Spates, Xinyue Ye, and Ashley Johnson. 2018. "i just might kill myself": Suicide expressions on twitter. *Death studies*.

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*, 170:155–160.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2020. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1):7.

Nicholas Tarrier, Patricia Gooding, Lynsey Gregg, Judith Johnson, and Richard Drake. 2007. Suicide schema in schizophrenia: The effect of emotional reactivity, negative symptoms and schema elaboration. *Behaviour Research and Therapy*, 45(9):2090–2097.

Cornelis Van Heeringen and A Marušic. 2003. Understanding the suicidal brain. *The British Journal of Psychiatry*, 183(4):282–284.

V. Venek, S. Scherer, L. Morency, A. ". Rizzo, and J. Pestian. 2017. Adolescent suicidal risk assessment in clinician-patient interaction. *IEEE Transactions on Affective Computing*, 8(2):204–215.

Stefan Wojcik and Adam Hughes. 2019. Sizing up twitter users.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Henrik D Zachrisson, Kjetil Rödje, and Arnstein Mykletun. 2006. Utilization of health services in relation to mental health problems in adolescents: A population based survey. *BMC Public Health*, 6(1).

Xingshan Zeng, Jing Li, Lu Wang, Zhiming Mao, and Kam-Fai Wong. 2020. Dynamic online conversation recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3331–3341, Online. Association for Computational Linguistics.

Xingshan Zeng, Jing Li, Lu Wang, and Kam-Fai Wong. 2019. Joint effects of context and user history for

predicting online conversation re-entries. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2809–2818, Florence, Italy. Association for Computational Linguistics.

Xu Zhang, Yaxuan Ren, Jianing You, Chao Huang, Yongqiang Jiang, Min-Pei Lin, and Freedom Leung. 2017. Distinguishing pathways from negative emotions to suicide ideation and to suicide attempt: The differential mediating effects of nonsuicidal self-injury. *Journal of abnormal child psychology*, 45(8):1609–1619.

Michael Zimmer. 2009. Web search studies: Multidisciplinary perspectives on web search engines. In *International handbook of internet research*, pages 507–521. Springer.

Michael Zimmer. 2010. "but the data is already public": on the ethics of research in facebook. *Ethics and information technology*, 12(4):313–325.

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 the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.