A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media

Ramit Sawhney
Netaji Subhas Institute of Technology
ramits.co@nsit.net.in

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

Saumya Gandhi
Visvesvaraya National Institute of Technology
gandhisaumya8@gmail.com

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

Abstract

Social media's ubiquity fosters a space for users to exhibit suicidal thoughts outside of traditional clinical settings. Understanding the build-up of such ideation is critical for the identification of at-risk users and suicide prevention. Suicide ideation is often linked to a history of mental depression. The emotional spectrum of a user's historical activity on social media can be indicative of their mental state over time. In this work, we focus on identifying suicidal intent in English tweets by augmenting linguistic models with historical context. We propose STATENet, a time-aware transformer based model for preliminary screening of suicidal risk on social media. STATENet outperforms competitive methods, demonstrating the utility of emotional and temporal contextual cues for suicide risk assessment. We discuss the empirical, qualitative, practical, and ethical aspects of STATENet for suicide ideation detection.[1]

1 Introduction

Globally, close to 800,000 people die by suicide each year, and 20 times more people attempt suicide. Suicide is the second leading cause of death in the 15 to 29 year age group (WHO, 2014) with a rising suicide rate of 35% in the US since 1999 (Hedegaard et al., 2020). Extending clinical and psychological care to people showing suicidal ideation relies heavily on identifying those at risk. Tragically, 80% of patients do not undergo psychiatric treatment, and about 60% of those who died of suicide denied having suicidal thoughts to mental health practitioners (McHugh et al., 2019). Recent studies (Coppersmith et al., 2018) also show that people exhibiting suicidal ideation make frequent use of social media, e.g., Twitter, to share their mental state, with eight out of ten disclosing their suicidal thoughts and plans (Golden et al., 2009).

Figure 1: We study a user whose latest tweet is not indicative of suicidal intent. Without seeing the user's recent historic tweet, which shows self-harm tendencies, it is difficult to accurately assess suicidal risk. However, analyzing a user's tweeting history sequentially without factoring in time irregularities between tweets may lead to an inaccurate representation of a user's mental state. Time-aware modeling of the temporal dependency between historic tweets reduces the impact of tweets from 3 years ago, providing a more realistic risk assessment. *All examples in this paper have been paraphrased for user privacy (Chancellor et al., 2019).*

While recent advances in computational social science (Coppersmith et al., 2018; Ji et al., 2019) have made progress in assessing suicidal risk on social media, analyzing the linguistic traits of tweets is often not sufficient for accurate suicidal intent detection. Additional user-level contexts such as tweeting history can be instrumental in identifying a build-up of negative emotions that are often linked to suicide ideation (Oliffe et al., 2012; Robins et al., 1959). Such a build-up can occur weeks, months, or even years before the onset of suicidal ideation (Overholser, 2003) and suicidal activity can also be influenced by past ideation or suicide attempts (Van Heeringen and Marušic,

[1] https://github.com/midas-research/STATENet_Time_Aware_Suicide_Assessment

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7685–7697,
November 16–20, 2020.

2003). Analyzing the user history and emotion spectrum, as shown in Figure 1 can provide crucial context to estimate suicidal risk in a tweet authored by that user. Such an **Emotional Historic Context (EHC)** of a user over time can be characteristic of their mental health (Coppersmith et al., 2014).

Modeling temporal user context, either as a bag-of-tweets (Gaur et al., 2019), or sequentially (Cao et al., 2019; Matero et al., 2019) helps in identifying suicidal intent. However, in Figure 1, we show that the impact of varying time intervals between tweets is crucial for an accurate assessment. It is critical to model the large gap between the user's recent tweets that are collectively indicative of suicidal intent and those three years apart. Such uneven **Temporal Tweeting Irregularities (TTI)** ranging from seconds to years (Wojcik and Hughes, 2019) between successive tweets influence the assessment of a user's tweet differently. Sequential models such as Long Short Term Memory (LSTMs) networks assume that posting intervals are uniform, hindering the learning ability of a user's emotion spectrum over varying time intervals.

Contributions: Taking into account a user's emotional historic context and temporal tweeting irregularities, we propose STATENet: **S**uicidality assessment **T**ime-**A**ware **TE**mporal **Net**work, a neural framework that evaluates the presence of suicidal intent on social media **(Sec. 3.1)**. Building on transfer learning's success in Natural Language Processing, STATENet uses a dual transformer-based architecture to learn the linguistic and emotional cues in tweets. STATENet jointly learns from the language of the tweet **(Sec. 3.2)** to be assessed, and the historic Plutchik-based (Plutchik, 1980) emotional spectrum of a user in a time-sensitive manner **(Sec. 3.3)**. Through a series of experiments **(Sec. 4)** on real-world data **(Sec. 4.1)**, we show that STATENet significantly outperforms competitive methods **(Sec. 5)**, with the F1 Score of 80%. We demonstrate practical applicability through a qualitative analysis **(Sec. 5.4)**, and discuss the ethical implications of this study **(Sec. 6)**.

At a minimum, we establish validity for time-aware emotional temporal context for identifying suicide ideation on social media. 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 considered 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, STATENet would flag tweets as *"at-risk"* for suicidality as part of a human-in-the-loop system to support decisions about potential intervention.

2 Related Work

Traditional Methods: Researchers have developed various psychoclinical methods to measure suicidal risk (Pestian et al., 2016), such as the Suicide Probability Scale (Bagge and Osman, 1998), Depression Anxiety Stress Scales-21 (Crawford and Henry, 2003), Adult Suicide Ideation Questionnaire (wa Fu et al., 2007), Suicidal Affect-Behavior-Cognition Scale (Harris et al., 2015), etc. While these methods are professional and effective, they require participants to either answer questionnaires (Venek et al., 2017) or engage in interviews (Scherer et al., 2013), hence not reaching suicidal people who are either unable to access these resources or have a low motivation to seek professional help (Zachrisson et al., 2006; Essau, 2005). Studies suggest that taking a suicide assessment can negatively impact individuals showing depressive symptoms (Harris and Goh, 2016).

NLP Methods: In recent years, social media has shown promise in providing insights into the psychological state of individuals (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 include the use of user features (Masuda et al., 2013) and online suicide notes (Pestian et al., 2010; Huang et al., 2007). Since then, the focus has been on using psycholinguistic lexicons such as LIWC (De Choudhury et al., 2016; Sawhney et al., 2018b) and textual features such as POS, tense, etc. for classification (Ji et al., 2018; Huang et al., 2014). Shared tasks such as CLPsych (Zirikly et al., 2019) and CLEF eRISK (Losada et al., 2019) have seen a rise in the use of deep learning for suicidality prediction. CNN based architectures (Du et al., 2018; Sawhney et al., 2018a; Shing et al., 2018; Naderi et al., 2019) and LSTM based architectures (Ji et al., 2018; Tadesse et al., 2020) utilize pre-trained word embeddings to predict suicide risk. Although these text-based methods capture the semantic nature of posts in isolation, no user associated context is provided

[2]Similar to the type of algorithmic model deployed for post level screening on Facebook (Card, 2018).

that can give insight into the user's mental state to improve predictive power (Venek et al., 2017). A user-dependent, personalized context can truly process the "natural" language of a user and understand the semantic context from the perspective of that specific user (Flek, 2020). User context may include the user's emotion spectrum (Ren et al., 2016), social graph methods (Mishra et al., 2019) and temporal context (Mathur et al., 2020). Suicide risk assessment for preliminary screening has been done at both binary (suicidal intent present, suicidal intent absent) (Cao et al., 2019; De Choudhury et al., 2016; Mathur et al., 2020; Losada et al., 2019), and multiple (Zirikly et al., 2019; Vioules et al., 2018; Gaur et al., 2019) levels of risk ranging from no risk to severe risk.

Contextual Methods: The best performing model, the dual context BERT (Matero et al., 2019), at the CLPsych 2019 shared task (Zirikly et al., 2019) for suicidal estimation on Reddit exemplifies the utility of temporal context. The Dual Context BERT utilizes post level BERT embeddings passed sequentially through an attention-based RNN. Similarly, Cao et al. (2019) employ a LSTM and fastText-based architecture for modeling temporal context. These RNN and LSTM based approaches assume that users' historical posts are equally spaced in time, hindering the suicide ideation detection model's ability to learn their relative importance in a time-aware manner. Time-aware sequential models have shown improvements in other clinical tasks (Baytas et al., 2017), such as patient subtyping, and in other domains like user activity modeling (Zhu et al.). More recently, Mathur et al. (2020) and Sinha et al. (2019) have modeled a user's historic emotion spectrum using latent representations of GloVe embeddings of historic tweets. These latent features are then aggregated based on specific functions such as exponential decay and sinusoids as opposed to learning them as sequences. These approaches assume that suicidal ideation conforms to specific trajectories, which may not generalize well across users (Giletta et al., 2015) and lose the context of individual historic tweets by aggregating them. Approaches besides deep learning have also been explored, such as the work done by Vioules et al. (2018), which uses the martingale framework (Ho, 2005) with sentiment scores and tweet level features such as likes to study two users on Twitter.

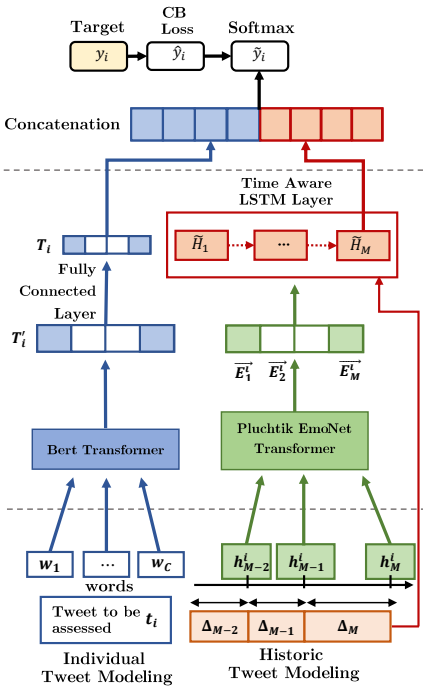


Figure 2: STATENet: Model Architecture

3 Methodology

3.1 Notations and Problem Formulation

We acknowledge that modeling suicidal intent as a binary classification task is a strong simplification and in this work, we focus on identifying the presence of suicide ideation within a tweet using a user-level temporal context. We denote a tweet to be assessed for suicidal risk as $t_i \in T = \{t_1, t_2, \cdots, t_N\}$ authored by a user $u_j \in U = \{u_1, u_2, \cdots, u_M\}$ made at time τ^i_{curr}. Each tweet t_i is associated with history $H_{i,j} = [(h^i_1, \tau^i_1), (h^i_2, \tau^i_2), \cdots, (h^i_L, \tau^i_L)]$ where h^i_k is a historic tweet by the user u_j posted at time τ^i_k with $\tau^i_1 < \tau^i_2 < \cdots < \tau^i_L < \tau^i_{curr}$. We formulate the problem as a classification task to predict a label y_i for the tweet t_i, where, $y_i \in$ {suicidal intent present, suicidal intent absent}.

3.2 Encoding the Tweet to be Assessed

Studies have shown that the linguistic styles of social media users can aid in understanding their mental state (De Choudhury et al., 2013) and that their suicidal behaviour is correlated with suicidal tweets (Sueki, 2015). Static word embeddings such as GloVe (Pennington et al., 2014) have been used to encode tweets for detecting suicide ideation (Sinha et al., 2019) in the past. However, recent

studies have shown that pre-trained transformer models yield more comprehensive representations of linguistic features in a tweet (Salminen et al., 2020). We found that SentenceBERT (Reimers and Gurevych, 2019) empirically outperforms embeddings used in previous works such as FastText (Cao et al., 2019), ELMo (Mohammadi et al., 2019), etc. We use the 768-dimensional encoding obtained from SentenceBERT.[3] Formally,

$$T_i' = \text{SentenceBERT}(t_i) \quad (1)$$

where $T_i' \in \mathbb{R}^{768}$ is linearly transformed using a dense layer to $T_i \in \mathbb{R}^d$ with dimension d.

3.3 User Historical Emotion Spectrum

Individual Historic Tweet Encoding: Amplification of emotional factors such as emotional reactivity (Tarrier et al., 2007), intensity (Links et al., 2008) and instability (Palmier-Claus et al., 2012) can increase suicide risk. Building on this, we extract the emotion spectrum of each historic tweet h_k^i. Although proficient in semantic modeling of text, general text encoders fail to capture the fine-grained emotions expressed in social media posts. To capture fine-grained emotions, we utilize Plutchik's wheel of emotions (Plutchik, 1980). This taxonomy suggests three hierarchical sets of eight emotions arranged as four pairs of opposing dualities. The primary set of emotions described by the wheel are: Joy - Sadness, Surprise - Anticipation, Anger - Fear, and Trust - Disgust. We obtain an encoding that models the emotional spectrum of a historical tweet, and thus that of a user at a historic time. Based on empirical comparisons and the success of transfer learning in NLP, we fine-tune pre-trained BERT embeddings on the Emonet dataset (Abdul-Mageed and Ungar, 2017). The dataset consists of a total of 1,608,233 tweets labeled across 24 emotions as per Plutchik's wheel of emotions. The presence of the primary emotions in the dataset is skewed towards joy, sadness, and fear, with their representation being 20.57%,8.85%, and 6.13%, respectively, with other emotions having fewer samples. These are labeled using distant supervision using a total of 665 emotion hashtags.

We call this transformer the PlutchikTransformer. This transformer tokenizes each historical post and adds the `[CLS]` token at the beginning of each post. We use the final hidden state corresponding to this

[3]SentenceBERT computes the mean of output vectors for all tokens to derive a fixed size sentence embedding.

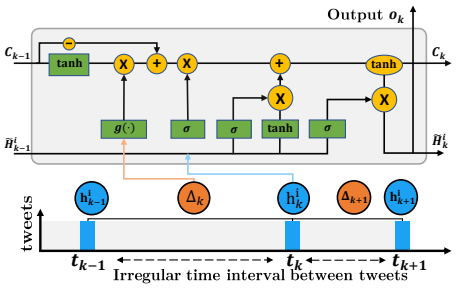


Figure 3: Architecture of a Time-aware LSTM cell. Figure is adapted from Baytas et al. (2017).

`[CLS]` token (768-dimensional encoding) as the aggregate representation of the emotional spectrum. We define the emotion vector ($E_k^i \in \mathbb{R}^{768}$) of each historic tweet h_k^i as:

$$E_k^i = \text{PlutchikTransformer}(h_k^i) \quad (2)$$

Modeling Historical Tweets Sequentially: The emotional historic context of tweets can be used to model progressive emotional states of the author of those tweets (Abdul-Mageed and Ungar, 2017; De Choudhury et al., 2013). This makes recurrent neural networks (RNN), and particularly LSTMs (Hochreiter and Schmidhuber, 1997), the most natural methods for encoding and learning from a sequence of a user's historical tweets.
However, the time interval between the posting of historic tweets can vary widely, from a few seconds to a few years (Wojcik and Hughes, 2019). Such variations can be an important factor in analyzing the emotional states of a user over time (Sueki, 2015). LSTM cells assume the input to be equally spaced sequences and thus are unable to model irregularities in posting times of historical tweets. Using this relative time difference between the user's historical tweets can progressively model the user's emotions more accurately over time. Hence, we propose the use of a Time-aware LSTM (T-LSTM) (Baytas et al., 2017) where time lapse between successive tweets is fed to the T-LSTM cell, as shown in Figure 3. The T-LSTM cell thus incorporates the actual time differences between tweets, along with each historical tweet's emotional context E_k^i.

T-LSTM applies time decay to the memory according to the elapsed time between successive elements and weights the short-term memory cell C_k^S. Intuitively, the greater the time elapsed be-

tween two tweets, the less impact they should have on each other. To achieve this, T-LSTM uses a monotonically decreasing function of elapsed time, which transforms time into appropriate weights. Time lapses are incorporated in the T-LSTM as:

$$
\begin{aligned}
C^S_{k-1} &= \tanh(W_d C_{k-1} + b_d) && \text{(Short-term memory)} \\
\hat{C}^S_{k-1} &= C^S_{k-1} * g(\Delta_k) && \text{(Discounted short-term memory)} \\
C^{LT}_{k-1} &= C_{k-1} + C^S_{k-1} && \text{(Long-term memory)} \\
C^*_{k-1} &= C^{LT}_{k-1} + \hat{C}^S_{k-1} && \text{(Adjusted previous memory)}
\end{aligned}
$$

where C_{k-1} and C_k are previous and current cell memories, and $\{W_d, b_d\}$ are network parameters. Δ_k is the elapsed time between historic tweets h_{k-1} and h_k, and $g(\cdot)$ is a heuristic decaying function that reduces the effect of short-term memory as Δ_k increases. We select $g(\Delta_k) = 1/\Delta_k$ empirically and as suggested in Baytas et al. (2017). For each historic tweet h^i_k, the T-LSTM cell modifies LSTM gate operations to compute the current hidden state ($\tilde{H}^i_k \in \mathbb{R}^d$) by feeding C^*_{k-1} instead of C_{k-1}.

3.4 Joint Network Optimization

To identify the presence of suicidal intent in a tweet, STATENet jointly learns from the language of the tweet to be assessed and the emotional historic spectrum in a time-aware manner. For this we apply the concatenation operation $\oplus$ to T_i and $\tilde{H}^i_k$ respectively, followed by a dense layer with Rectified Linear Unit ($ReLU$) (Hahnloser et al., 2000) to form a prediction vector. Finally, a softmax function (Goodfellow et al., 2016) is used to output the probabilities of suicidal intent present.

$$
\begin{aligned}
\tilde{y}_i &= ReLU(W_y(T_i \oplus \tilde{H}^i_k) + b_y) \\
\hat{y}_i &= softmax(\tilde{y}_i)
\end{aligned}
\tag{3}
$$

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

Tweet indicating suicidal intent form a very small proportion of the data (Ji et al., 2019). To address this problem of class imbalance *(in practice, the imbalance is much greater in the real world)*, we train STATENet using Class-Balanced loss proposed by Cui et al. (2019) along with Focal Loss (Lin et al., 2017). This loss function applies a class-wise re-weighting scheme by introducing a weighting factor that is inversely proportional to the number of samples. The loss function $\mathcal{L}$ is:

$$
\mathcal{L} = \text{CB}_{focal}(\hat{y}_i, y_i; \beta, \gamma) \tag{4}
$$

where CB_{focal} is class-balanced focal loss, $\hat{y}_i$ is the predicted label and y_i is the label of the current tweet. β and γ are hyperparameters.

4 Experiments

4.1 Dataset

We use the Twitter timeline data of users from the dataset introduced by Sinha et al. (2019). Sinha et al. (2019) began with a collection of Twitter posts based on a lexicon of 143 suicidal phrases. After manual inspection of the dataset for trivially non-suicidal tweets, their final dataset contained 34,306 tweets. Some of these tweets were authored by the same user; thus, the total number of unique users for which tweets were to be classified was 32,558. We summarize the annotation instructions (Sawhney et al., 2018b) that were followed by two annotators, both students of Clinical Psychology, for annotating the collected 34,306 tweets:

- **Suicidal Intent (SI) Present**: Posts 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, including song lyrics, condolence message, awareness, news.

It is important to note that this process produced suicide risk labels at the level of individual tweets and not for individual user histories. An acceptable inter-annotator agreement was achieved with a Cohen's Kappa score (Cantor, 1996) of 0.72, under the supervision of a professional clinical psychologist. The resulting dataset contains 3984 suicidal tweets. The Twitter timeline was collected for each user. These timelines span over ten years from 2009 to 2019. The mean number of tweets in user history is 748 *(max 3,200)* with a standard deviation of 789 tweets. We trim the user history to the 100 most recent tweets for users with a large number of historical tweets.[4] The mean time difference between two consecutive tweets for a user is two days with a standard deviation of almost 24 days between two tweets, indicative of large variations across users. 4070 users were found to have no historical tweets.

Data Preprocessing: We deidentified the dataset by performing named entity recognition and removing any identifiable information such as email

[4]This was done due to memory and computation constraints faced during the training of STATENet.

addresses, URLs, and names. Next, we follow standard procedures of converting the text to lowercase, removing punctuation and accents, striping whitespaces, and removing stopwords. We split the tweets in the dataset on the basis of users such that there is no overlap between users in the train, validation, and test set. We perform a stratified 70:10:20 split across the three sets, such that the train, validation, and test sets consist of 24014, 3431, and 6861 tweets, respectively. Although there may be multiple tweets to be assessed by the same user, their associated history differs according to the tweets' posting timestamps. We ensure that for each tweet to be classified, only the historical tweets having timestamps older than that of the tweet to be assessed are used for historic modeling.

4.2 Experimental Settings

Baseline Methods: We evaluate STATENet using the macro F1 and recall for *suicidal intent present* (recall^s), against two types of baseline methods; tweet level (TL) and user-level (UL). UL baselines were adapted for tweet level assessment by concatenating embeddings of the tweet to be assessed with the user level features.

Random Forest + Tweet features (Sawhney et al., 2018b): A non contextual TL approach that applies Random Forests (RF) with tweet level features including statistical, LIWC (Pennebaker et al., 2001) features, n-grams and POS counts.

C-LSTM (Sawhney et al., 2018a): We replicate the TL deep Neural Network that uses CNN to capture local features and LSTMs for tweet encoding.

Suicide Detection Model (SDM) (Cao et al., 2019): UL model that encodes tweets using fine-tuned FastText embeddings. Historic tweets were passed sequentially through LSTM + attention and concatenated with the tweet to be assessed.

Contextual CNN (Gaur et al., 2019): Non-sequential UL model using GloVe embeddings for encoding tweets. Bag of tweets were concatenated and fed to a contextual CNN (Shin et al., 2018).

Exponential Decay (Sinha et al., 2019): TL model that weighs GloVe embeddings of historic tweets through an exponential decay function and ensembles it with the GloVe embedding trained on a BiLSTM + Attention for the tweet to be assessed.

Surprise and Episodic Modeling (Mathur et al., 2020): Decision level ensemble TL model similar to Exponential Decay, but factors in sinusoidal and white Gaussian noise for historic tweet modeling.

DualContextBert (Matero et al., 2019): Best performing UL model at CLPsych 2019. DualContextBert uses BERT for encoding Reddit posts fed to an attention-based RNN layer. In our implementation, we use all the user's historic tweets.

Experimental Setup: We select hyperparameters based on the highest Macro F1 obtained on the validation set for all models. We use grid search to explore: number of features in hidden state $\tilde{H}^D \in \{8, 64, 128, 256, 512\}$, number of LSTM layers $n \in \{1, 2, 5\}$, dropout $\delta \in \{0.0, 0.1, \cdots, 0.5\}$, $\beta \in \{0.99, 0.999, 0.9999\}$ and $\gamma \in \{1.0, 1.5, 2.0\}$ in class-balanced focal loss, initial learning rate $I_{lr} \in \{0.01, 0.001, 0.0005, 0.0001\}$, warm-up steps $S_{ws} \in \{3, 5, 7\}$. The optimal hyperparameters were found to be: $\tilde{H}^D = 512$, $n = 1$, $\delta = 0.5$, $\beta = 0.9999$, $\gamma = 2.0$, $I_{lr} = 0.0001$, $S_{ws} = 5$. We implement all methods with PyTorch 1.5 (Paszke et al., 2019) and optimize using mini-batch AdamW with a batch size of 256 and $I_{lr} = 0.0001$. We use the cosine scheduler with a warmup step of 5 (Gotmare et al., 2018). We train the model for 20 epochs and apply early stopping with a patience of 5 epochs. The model takes 4,361s to train on an Nvidia Tesla K80 GPU.

5 Results and Analysis

5.1 Comparative Performance

We note from Table 1 that STATENet significantly ($p < 0.005$) outperforms competitive baselines. We compare against both text only, and temporal contextual models for suicidal risk assessment.

STATENet and other contextual models perform better than the non-contextual RF + tweet features and C-LSTM models. We believe this is because temporal contextual models offer greater insight into the author's historical mental state, thereby increasing predictive power. STATENet and sequential models outperform the Contextual CNN, likely due to their ability to better learn representations from the temporal dependence in historical tweets, as opposed to Contextual CNN's bag of tweets approach. We also observe that STATENet significantly outperforms competitive sequential models. We postulate this to the ability of the time-aware LSTM in STATENet to capture irregularities in tweeting intervals of users. Such time-aware modeling likely learns more accurate latent representations of users' emotional historic context. While exponential decay and episodic modeling

Type of Contextual Modeling	Model	Macro F1 ↑	Recall^s ↑	Accuracy ↑
None	Random Forest + Tweet features	0.536	0.513	0.548
	C-LSTM	0.588	0.597	0.602
Non Sequential	Contextual CNN	0.729	0.587	0.803
Sequential	Suicide Detection Model (SDM)	0.743	0.755	0.819
	DualContextBert	0.767	0.786	0.823
Specific Temporal Functions	Exponential Decay	0.737	0.759	0.828
	Surprise and Episodic Modeling	0.741	0.762	0.831
Timeaware Sequential	**STATENet**	**0.799***	**0.810***	**0.851***

Table 1: Mean of results obtained over 10 different runs. * indicates that the result is significantly better than DualContextBert ($p < 0.005$) under Wilcoxon's Signed Rank test). **Bold** denotes best performance.

Model Component	Macro F1↑	Recall^s↑
Current tweet only	0.731	0.551
Current + Random History *(Plutchik)*	0.730	0.608*
Current + Sequential History *(BERT)*	0.767*	0.786*
Current + Sequential History *(Plutchik)*	0.778*	0.795*
Current + TA History *(Plutchik)*	**0.799***	**0.810***

Table 2: Ablation study over STATENet. We report the mean of results obtained over ten different runs. * shows significant compared to the current tweet ($p < 0.005$) under Wilcoxon's Signed Rank test. Current: encoding of the tweet to be assessed. History: encoding of historical tweets. TA: Time-Aware. **Bold** denotes the best performance.

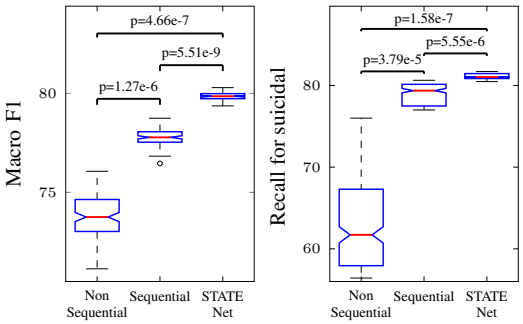


Figure 4: Confidence intervals for evaluation metrics of temporal variants over 10 different runs and data splits.

perform well, we note that STATENet does better, in terms of all metrics, particularly recall for the suicidal intent present class. We believe this is because not every user's emotional historic context may conform to fixed trajectories that these approaches aggregate historic tweets on.

5.2 Ablation Study

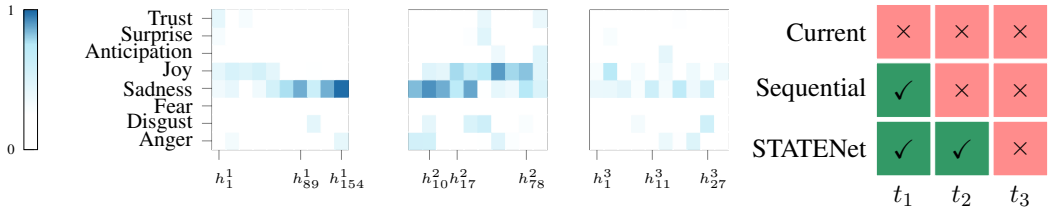
To assess EHC and TTI, we perform an ablation study (Table 2) with different configurations. Without considering historic tweets, the performance of the model drops drastically. We believe that adding historic tweets, even in a random order, adds additional contextual cues about the user, resulting in improved performance. We observe that the *PlutchikTransformer* variant of Current + Sequential History outperforms its *BERT* counterpart. This can be attributed to the ability of the *PlutchikTransformer* to capture the EHC of a user. STATENet jointly models the Current Tweet and EHC in a time-aware manner, overcoming the limitation of previous models that assume equal time intervals between posts. On inspecting the results for the 647 users without any historic tweets, we find that STATENet performs well with a recall of 0.74 and macro F1 of 0.75. This reiterates the ability of linguistic only non-contextual models in suicidal intent identification. This is particularly interesting, as, for users with no available history, assessment can still be performed to some degree.

5.3 Temporal Analysis

The tweet's language should be studied with historical context to better understand the user's emotional state over time, based on the EHC. To analyze the importance of the order and temporal dependency of historic tweets, we first try a non-sequential, bag of tweets like variant. We feed the Plutchik transformer-based encodings to a Contextual CNN. We observe that the bag of tweets approach is slightly better than the Contextual CNN baseline, likely because of the transformer-based encoding as opposed to static GloVe embeddings used by the baseline. The non-sequential approach drastically underperforms over ten runs in comparison to temporal variants. Further investigating EHC and TTI, we first feed historic tweets in sequential order (Sequential Model) to a regular LSTM, and then we factor in TTI through T-LSTM in STATENet. Figure 4 shows that STATENet is

	User 1	User 2	User 3
t_i	[14/07/2017]: i dont want to be here anymore again *(SI Present)*	[08/03/2016]: been a year since i lost the most important woman the loss has never sunk *(SI Absent)*	[05/01/2019]: Nobody be shocked when I snap and take a life either my own or theirs *(SI Present)*
$h^i_{k_3}$	[13/07/2017]: yes i almost tried to kill myself again tonight yes it is only been ten minutes and im now retweeting tweets	[16/11/2015]: i wrote this a year ago today and one year on i am boxing things up and moving into my own house it is crazy	[29/12/2018] when you said your last goodbye, i die a little bit inside, i lay in tears in bed all night, alone without you by my side
$h^i_{k_2}$	[27/05/2017]: i do not know if its seasonal depression or just me avoiding christmas by staying in all day	[19/11/2014]: I don't think i will see the end of today, there is nothing left for me to do	[21/12/2018]: Do you collect anything. If so what is it? Memories hahhahahhaa
$h^i_{k_1}$	[02/06/2016]: i love my mother she is great life is amazing	[16/11/2014]: i deserve death, dear 16 old me it will never get better	[16/12/2018]: I am alive and I am happy about it dammit why even

Table 3: Tweet to be assessed (t_i) and historic tweets ($h^i_{k_1}$, $h^i_{k_2}$ and $h^i_{k_3}$ are chronologically ordered) of three users along with tweet timestamp information. We also show visualized self-attention (averaged over all 12 Sentence-BERT attention heads) per token. Darker intensity of the red color denotes higher attention weights.



(a) User 1 (b) User 2 (c) User 3 (d) Model predictions for tweets

Figure 5: (a), (b) and (c) are emotion intensity across 8 primary emotions based on the Plutchik Wheel for User 1, 2, 3 over time respectively, from White to Blue. h^i_k represents k^{th} historic tweet associated with the current tweet t_i. In (d) Green and Red represent correct and incorrect assessment of suicidal risk respectively for different tweets. We display only the 8 primary emotions of the Plutchik wheel for brevity.

significantly ($p < 0.005$) better than the sequential but non-time-aware variant, and shows the least variation in performance over 10 different runs. For the difference in performance between the Sequential Model and STATENet, we believe that is due to the temporal dependency of historic tweets on the elapsed time between successive tweets.

5.4 Qualitative Analysis

For a detailed insight and aiding interpretability, we analyze some cases where STATENet performs well. We also highlight the limitations of STATENet through error analysis. We qualitatively analyze three interesting cases in Table 3 and Figure 5. We see that the tweet to be assessed for User 1 does not show any explicit suicidal intent and alone may not be sufficient to assess suicidal risk. However, temporal models correctly classify the tweet as they learn the build-up of sadness in the historic tweets, which we observe from the Pluchtik emotional intensity in Figure 5a.

When the current tweet of the user is non-indicative, temporal models can get additional context by learning historic activity of the user. Often, temporal patterns are variable, and posting frequencies vary drastically. These TTI present challenges in only relying on the sequence of historic tweets rather than the actual time lapses. For instance, initial tweets of User 2 showed sadness and suicidal intent, whereas the recent historic tweet ($h^2_{k_3}$) of the user represents joy (Figure 5b). LSTM-based models aggregate sadness and hence assume the history to be suicidal. Contrarily, STATENet is able to learn from the variable time-lapses and their relative importance in the context of suicide ideation. However, we found some cases where all models failed. For User 3, the current tweet does not contain strong semantic indicators of suicidal intent. Moreover, historic tweets do not show any recognizable emotional pattern (Figure 5c). Such a case presents the complexities associated with suicide risk assessment.Another interesting observation from Figure 5 is that the learned Plutchik emotion

intensity distribution for users is skewed towards joy (positive) and sadness (negative). Although the highly granular emotional context captured by the PlutchikTransformer improves STATENet's performance (**Sec.** 5.2), over the more generic language features captured by BERT. We leave further exploring the impact of emotion granularity to our future research directions.

6 Discussion

Ethical Considerations: The preponderance of the work presented in our discussion presents heightened ethical challenges. As explored in Coppersmith et al. (2018), we address the trade-off between privacy and effectiveness. While data is essential in making models like STATENet effective, we must work within the purview of acceptable privacy practices to avoid coercion and intrusive treatment. To that end, we utilize publicly available Twitter data in a purely observational (Norval and Henderson, 2017; Broer, 2020), and non-intrusive manner. Although informed consent of each user was not sought as it may be deemed coercive, automated de-identification of the dataset was performed to reduce the risk of including any identifying data in the raw data. All tweets shown as examples in Figure 1 and Section 5.4 have been paraphrased as per the *moderate disguise* scheme suggested in Bruckman (2002) to protect the privacy of individuals (Fiesler and Proferes, 2018). The annotation of user data has been kept separately from raw user data on protected servers linked only through anonymous IDs (Benton et al., 2017). Assessments made by STATENet are sensitive and should be shared selectively to avoid misuse, such as Samaritan's Radar (Hsin et al., 2016). Our work does not make any diagnostic claims related to suicide. We study the social media posts in a purely observational capacity (Norval and Henderson, 2017) and do not intervene with the user experience in any way.

Limitations: We acknowledge that studying suicidality is subjective in nature (Keilp et al., 2012) and that the interpretation of the analysis presented may vary across individuals. Due to the situatedness of language, the studied data may be susceptible to demographic, annotator, and medium-specific biases (Hovy and Spruit, 2016). We recognize that suicide risk exists on a diverse spectrum, and the simplification of binary labels could lead to artificial notions of risk (Bryan and Rudd, 2006).

Practical Implications: Through STATENet, we suggest a neural architecture for preliminary screening of at-risk users on social media to aid the prioritization of clinical resources. Our work observes Twitter in a non-intrusive manner and does not intervene with the user experience in any way. STATENet should form part of a distributed human-in-the-loop (de Andrade et al., 2018) system for finer interpretation of risk. Focusing on STATENet's practical applicability, we work with tweet level annotations rather than the more subjective and difficult to scale user-level annotations. We emphasize on tweet-level prediction; however, STATENet can also be applied for user-level suicide risk assessment given its dual text and historic modeling components.

7 Conclusion

Motivated by the rising use of social media for exhibiting suicide ideation as opposed to standard clinical practice (McHugh et al., 2019), we present STATENet. Building on psychological studies on analyzing a user's temporal emotional spectrum, STATENet models the time aware emotional context of users through historical tweets for more accurate suicide risk estimation on social media. We plan to explore the impact of varying amounts of historical context for a user in our future work. We show STATENet's applicability as a preliminary tool in assessing suicidality in tweets. We present a qualitative analysis for a deeper understanding of STATENet. Through this work, we aim to form a component in a larger human-in-the-loop infrastructure for analyzing potentially concerning suicide-related social media posts. Priority-based suicide risk assessment for ranking tweets for suicidal risk, rather than classifying them forms our future direction. Additionally, in the future, we would also want to quantify the impact of varying degrees of granularity of learning emotional features from tweets on STATENet's performance.

Acknowledgements

We would like to think Alex Polozov, Kawin Ethayarajh, Sebastian Gehrmann, Siva Reddy, and the anonymous reviewers for their extremely helpful feedback and comments.

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.

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. 2017. 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.

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

Alan B Cantor. 1996. Sample-size calculations for cohen's kappa. *Psychological Methods*, 1(2):150.

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.

Catherine Card. 2018. How facebook ai helps suicide prevention. *Facebook Newsroom*.

Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. 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.

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.

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.

John R. Crawford and Julie D. Henry. 2003. The depression anxiety stress scales (DASS): Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology*, 42(2):111–131.

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.

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. 30066665[pmid].

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.

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, Online. Association for Computational Linguistics.

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.

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.

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

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*.

Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951.

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.

Shen-Shyang Ho. 2005. A martingale framework for concept change detection in time-varying data streams. In *Proceedings of the 22nd international conference on Machine learning*, pages 321–327.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

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.

Xiaolei Huang, Lei Zhang, Tianli Liu, David Chiu, Tingshao Zhu, and Xin Li. 2014. Detecting suicidal ideation in chinese microblogs with psychological lexicons. *CoRR*, abs/1411.0778.

Y. Huang, T. Goh, and C. L. Liew. 2007. Hunting suicide notes in web 2.0 - preliminary findings. In *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*, pages 517–521.

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.

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.

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. 2019. Overview of erisk at clef 2019 early risk prediction on the internet (extended overview).

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. *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*, 12036:265–271. PMC7148016[pmcid].

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.

Elham Mohammadi, Hessam Amini, and Leila Kosseim. 2019. CLaC at CLPsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 34–38, Minneapolis, Minnesota. Association for Computational Linguistics.

Nona Naderi, Douglas Teodoro, Emilie Pasche, and Patrick Ruch. 2019. A baseline approach for early detection of signs of anorexia and self-harm in reddit posts.

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

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*, 74(4):506–514.

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

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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*.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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. 21643548[pmid].

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Eli Robins, George E Murphy, Robert H Wilkinson Jr, Seymour Gassner, and Jack Kayes. 1959. Some clinical considerations in the prevention of suicide based

on a study of 134 successful suicides. *American Journal of Public Health and the Nations Health*, 49(7):888–899.

Joni Salminen, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences*, 10(1):1.

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.

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.

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*, CIKM '19, page 941–950, New York, NY, USA. Association for Computing Machinery.

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.

M. J. Vioules, B. Moulahi, J. Aze, and S. Bringay. 2018. Detection of suicide-related posts in twitter data streams. *IBM Journal of Research and Development*, 62(1):7:1–7:12.

WHO. 2014. *Preventing suicide: A global imperative*. World Health Organization.

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

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).

Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. What to do next: Modeling user behaviors by time-lstm.

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.