Suicide Ideation Detection via Social and Temporal User Representations using Hyperbolic Learning

Ramit Sawhney[†‡*] and **Harshit Joshi**[¶*] and **Rajiv Ratn Shah**[‡] and **Lucie Flek**[†]
† Conversational AI and Social Analytics (CAISA) Lab
Department of Mathematics and Computer Science, University of Marburg
`http://caisa-lab.github.io`
‡ Multimodal Digital Media Analysis (MIDAS) Lab
Department of Computer Science and Engineering, IIIT Delhi
`http://midas.iiitd.edu.in`
¶ Cluster Innovation Centre, University of Delhi

Abstract

Recent psychological studies indicate that individuals exhibiting suicidal ideation increasingly turn to social media rather than mental health practitioners. Personally contextualizing the buildup of such ideation is critical for accurate identification of users at risk. In this work, we propose a framework jointly leveraging a user's emotional history and social information from a user's neighborhood in a network to contextualize the interpretation of the latest tweet of a user on Twitter. Reflecting upon the scale-free nature of social network relationships, we propose the use of Hyperbolic Graph Convolution Networks, in combination with the Hawkes process to learn the historical emotional spectrum of a user in a time-sensitive manner. Our system significantly outperforms state-of-the-art methods on this task, showing the benefits of both socially and personally contextualized representations.

1 Introduction

Every 40 seconds, a person dies by suicide (Roth et al., 2018). Despite the success of psychoclinical methods, such as the Suicide Probability Scale (Bagge and Osman, 1998) and Suicide Ideation Questionnaire (wa Fu et al., 2007), the suicide rate in the U.S. has risen by 35% in the last 20 years (Hedegaard et al., 2020). While these methods are professional (Pestian et al., 2017), they have limited efficacy and may even impact participants negatively (Harris and Goh, 2017). Their limitations include barriers such as social stigma (Crisp et al., 2000), low literacy (Batterham et al., 2013), low motivation to seek hel (Essau, 2005), and finances (Czyz et al., 2013). Tragically, 80% of patients do not undergo clinical treatment, and 60% of those who died by suicide denied having suicidal thoughts to practitioners (McHugh et al., 2019).

Contrarily, people turn to social media to express suicidal thoughts (Luxton et al., 2012; Coppersmith et al., 2014; Robinson et al., 2016), with 8 of 10 people disclosing their suicidal plans (Golden et al., 2009). Consequently, a growing body of work has shown that natural language processing can complement social media analysis to identify risk markers in online user behavior to aid suicide risk assessment (McCarthy, 2010; De Choudhury et al., 2016; Reger et al., 2020; Shing et al., 2018). However, analyzing individual user posts is not always sufficient to infer user's mental state and the associated suicide risk (Harris, 2010; Sisask et al., 2008).

Studies suggest that suicide can be influenced by social factors (Masuda et al., 2013; Gvion and Apter, 2012), and is a contagious phenomenon (Mann, 2002). If a user is inclined to suicide ideation, a neighbor in the social network also often exhibits suicidal behavior (Wray et al., 2011). Further, social media cultivates safe spaces that encourage users to share thoughts with those who appear similar to themselves (Bak et al., 2012; McPherson et al., 2001; Franklin et al., 2017). Analyzing such social context along with historical activity, as in Figure 1, can help further ascertain suicidal risk (Van Heeringen and Marušic, 2003).

According to psychosocial research, there exists an uneven distribution of power and influence on social media (Avin et al., 2018). People exhibiting suicidal ideation form social clusters (Robertson et al., 2012) and preferentially copy the behavior of popular users, manifesting social learning of suicide-related behavior such as the "copycat suicide" (Mesoudi, 2009; Henrich and Gil-White, 2001) (Figure 1). These social networks present a hierarchical structure of ideation propagation, characteristic for **Scale-free** networks (Barabási and Bonabeau, 2003). In a scale-free network, most nodes have very few links, whereas a handful of in-

[*] Both authors contributed equally

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2176–2190
June 6–11, 2021.

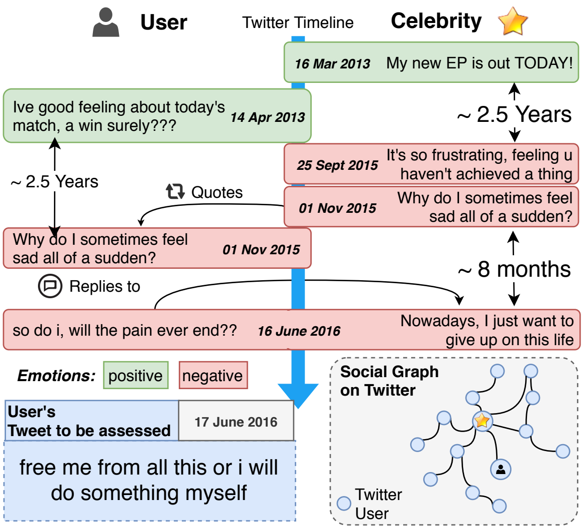


Figure 1: Illustration of social influence and context, specifically copycat suicidal ideation, in a scale-free network setting. Such social and temporal context can contextualize a user's state for a more accurate suicide risk assessment. We paraphrase all examples in this paper as per the moderate disguise scheme (Bruckman, 2002) to protect user privacy (Chancellor et al., 2019b).

fluential nodes have a large number of connections, creating social hubs, further amplifying phenomena such as the "Werther effect" (Fahey et al., 2018).

Social networks with scale-free structure are subjects to major distortions when embedded into the Euclidean representation space (Chen et al., 2013; Aparicio et al., 2015) by ordinary graph neural networks. To overcome this limitation, we propose to model the social relations using graph convolutions over hyperbolic space (Chami et al., 2019).

Our key contributions are as follows:

(i) We present the first deep graph neural framework to identify suicide ideation on social media by explicitly modeling users' social and temporal emotional context jointly (§3).

(ii) Motivated by psychological studies and the scale-free nature of social networks, we propose the use of Hyperbolic Graph Convolutions (§3.4).

(iii) We propose a mechanism leveraging Hawkes process to learn the historic emotional spectrum of a user in a time-sensitive manner from their historical posts (§3.3).

(iv) Through a series of experiments (§5), we show that our framework significantly outperforms existing methods (§6.1) on this task, as well as standard Graph Neural Networks (§6.2).

(v) Finally, we analyze the contributions of Hyper-SOS's individual components to assess suicidal intent (§6.2, §6.3, §6.4) and demonstrate practical applicability through a qualitative analysis (§6.5).

Aware of the sensitive nature of this work, we dedicate a standalone section (§7) to the ethical considerations and applicability of this work.

2 Related Work

2.1 Suicide Ideation Detection

Early efforts in leveraging NLP for suicide ideation detection on social media (De Choudhury et al., 2013, 2016; Shing et al., 2018; Sawhney et al., 2018) combine general features such as n-grams and POS tags with lexicons like LIWC (Pennebaker et al., 2001). Deep learning models like CNNs (Naderi et al., 2019) and LSTMs (Coppersmith et al., 2018) have improved suicide ideation detection (Ji et al., 2020) thanks to a more robust semantic context to interpret the tweet in question, however, lacking user-level context, are often unable to ascertain suicide risk (Sisask et al., 2008). The best performing models (Matero et al., 2019; Naderi et al., 2019) at the CLPsych (Zirikly et al., 2019) and CLEF e-Risk (Losada et al., 2019) exemplify the promising yet underexplored direction of user context modeling (Flek, 2020) for suicide ideation detection. Although recent studies (Shing et al., 2020; Sawhney et al., 2020) explore the personal historical context of users, community-based social context has rarely been explored for this task. One of the few attempts includes SNAPBATNET (Sinha et al., 2019), a shallow embedding model to extract network structural features.

2.2 Graph Neural Networks

While graph neural networks (GNNs) have made advances in enhancing NLP models for various tasks (Mishra et al., 2019a; Del Tredici et al., 2019; Lu and Li, 2020), two broad shortcomings limit their effectiveness for suicide ideation detection. First, these methods do not capture the personal historical and social network context together, both of which are strongly correlated to risk assessment on social media (Yang and Eisenstein, 2017). Second, studies have shown that users exhibiting suicide ideation tend to form social networks with scale-free characteristics (Jonas, 1992; Mesoudi, 2009), which regular GNNs are unable to accurately capture (Chami et al., 2019) in learnt social representations.We build on these limitations by combining historical and social contexts in the hy-

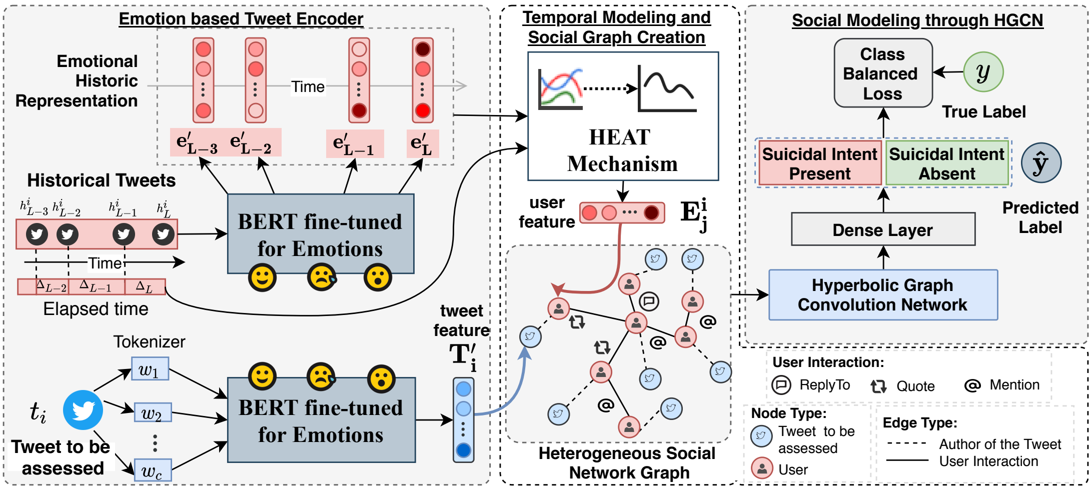


Figure 2: An overview of Hyper-SOS: We first extract the emotional representation of the tweet to be assessed and the historic emotional spectrum of a user via the HEAT mechanism to initialize tweet nodes and user nodes in the heterogeneous social graph, respectively. A Hyperbolic GCN is then used to aggregate features from neighboring nodes to learn social and historic representation, which we use to assess the presence of suicidal intent.

perbolic space to further contextualize and improve suicide ideation detection on social media.

3 Hyper-SOS: Formulation and Design

In this section we present the architecture of the Hyper-SOS framework (**Hyper**bolic Graph Convolutional Network for **S**uicide assessment **O**n **S**ocial media) shown in Figure 3, designed to identify suicide ideation on social media by explicitly modeling user's social and temporal emotional context.

3.1 Problem Formulation

We formulate suicidal intent (SI) detection as a binary classification task to predict the presence of suicidal intent y_i for a tweet t_i, where, $y_i \in \{\text{SI present, SI absent}\}$. 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 $\tau^i_1 < \tau^i_2 < \cdots < \tau^i_L < \tau^i_{curr}$. Moreover, two users are connected if they interact with each other's tweets on Twitter. We acknowledge that modeling suicidal intent as a binary classification task is a strong simplification.

3.2 Encoding Tweets

We build on previous studies which show that the linguistic styles (De Choudhury et al., 2013, 2016) and emotions expressed in suicidal tweets play an important role in assessing suicidal behavior (Sueki, 2015; Zhang et al., 2017; Spates et al., 2018). Thus, building on this correlation between emotions and suicidal ideation, we fine-tune BERT on EmoNet (Abdul-Mageed and Ungar, 2017) for capturing fine-grained (Plutchik-based) emotions (Plutchik, 1980; Sawhney et al., 2020).

Tweet to be assessed: We utilize the final 768-dimension hidden state corresponding to the `[CLS]` token as the aggregate representation of emotions in a tweet. Formally, we encode each tweet to be assessed (t_i) to an emotion representation vector $\mathbf{T'_i} = \text{BERT}_{\text{finetuned}}(t_i)$; $\mathbf{T'_i} \in \mathbb{R}^{768}$.

Historical Tweets: We encode user's historical tweets h^i_k using our fine-tuned BERT to learn representations of a user's emotional spectrum over time as $\mathbf{e^i_k} = \text{BERT}_{\text{finetuned}}(h^i_k)$, $\mathbf{e^i_k} \in \mathbb{R}^{768}$. These representations can be indicative of a user's mental state and emotion buildup over time (Aragón et al., 2019; Tarrier et al., 2007), and better contextualize temporal behavior to ascertain suicidal intent (Links et al., 2008; Palmier-Claus et al., 2012).

3.3 Modeling Personal Historical Context

To model historical emotions of a user and factor in the natural irregularities in posting time of historical tweets (Lei et al., 2018; Wojcik and Hughes, 2019), we propose the HEAT mechanism: **H**awkes temporal **E**motion **A**ggregra**T**ion. HEAT leverages Hawkes Process (Hawkes, 1971), a self-exciting temporal point process to model the intensity of emotions whenever a tweet is posted in the past (Guo et al., 2019). Intuitively, it assumes that emotions exhibited in different historic tweets can influence one another. To obtain the final historic representation ($\mathbf{E^i_j} \in \mathbb{R}^{768}$) of the tweet to be assessed t_i, HEAT aggregates encoded historical emotions $\mathbf{e^i_k}$ using an exponential kernel as:

$$\mathbf{E^i_j} = \sum_{k:\Delta\tau_k \geq 0} (\mathbf{e^i_k} + \epsilon \mathbf{e^{i'}_k} e^{-\beta\Delta\tau_k}),\ \mathbf{e^{i'}_k} = \max(\mathbf{e^i_k}, 0) \quad (1)$$

where, $\Delta\tau_k$ is the time gap between a historical tweet and the tweet to be assessed (current tweet) posted at time τ_k and τ_{curr}, respectively. ϵ and β are hyperparameters such that $\epsilon < \beta$.

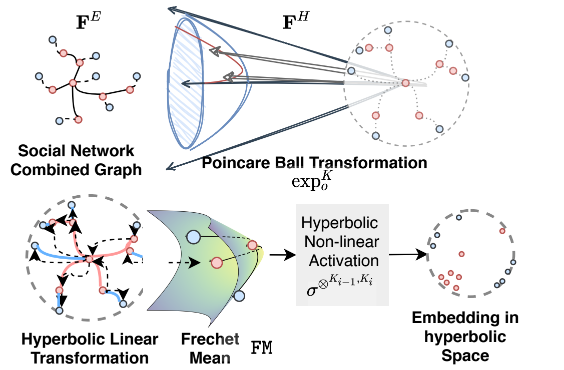


Figure 3: Hyperbolic feature transformation ($F^E \rightarrow F^H$) via projection on the Poincaré ball manifold to better represent the scale-free social network (left). Neighborhood-based node feature updation via hyperbolic linear transformation followed by Frechet Mean aggregation (right) to enrich user and tweet features.

3.4 Modeling Social Network Context

Studies show that users' emotions (Hill et al., 2010, 2015), depressive behavior (Rosenquist, 2011), and loneliness (Cacioppo et al., 2009) can be transmitted through social connections. Hence, leveraging social relationships between users can contextualize potential suicidal intent (Mueller and Abrutyn, 2015; Burnap et al., 2015; Colombo et al., 2016).

We model such relationships as a graph $\mathcal{G} = (V, E)$, where each edge $e^U \in E$ represents one of three types of interaction between two users $u_x, u_y \in U$: i) User u_x **quotes** (retweets) a tweet t_i, posted by user u_y, ii) User u_x **mentions** user u_y in a tweet t_i, iii) User u_x **replies** to user u_y, by posting a tweet t_i. We further extend the social graph $\mathcal{G}$ by introducing tweet nodes $t \in T$, which represent labeled tweets to be assessed. Each tweet node t is connected to its author (user) node u by a user-tweet interaction edge $e^T \in E$. The constructed social graph $\mathcal{G}$ is heterogeneous, having two types (users and tweets) of nodes $V = \{U \cup T\}$, and two types (user-user and user-tweet) of edges $E = \{e^T \cup e^U\}$, as shown in Figure 2. Note that the tweet nodes t are labeled for the presence of suicidal intent, while the user nodes u are unlabeled.

3.5 Hyperbolic Graph Neural Network

To augment language and historical context-based features, we leverage GNNs to learn representations of the constructed social graph $\mathcal{G}$. However, most GNNs such as Graph Convolution Networks (GCNs) operate in the Euclidean space, and often do not generalize well to the kind of hierarchical, tree-like networks users on social media, particularly those exhibiting suicidal behavior (Chen et al., 2013). Sociological studies (Bild et al., 2015; Aparicio et al., 2015), show that such networks show **scale-free** characteristics (Scatà et al., 2018), which follow the power law, i.e., the degree distribution of nodes decreases exponentially with a few nodes having a large number of connections (Ravasz and Barabási, 2003). To capture such hierarchical and scale-free structural properties in the social network graph, we propose the use of a Hyperbolic Graph Convolution Network (HGCN) (Chami et al., 2019). HGCNs project language and historical feature embeddings in the hyperbolic space to minimize distortions and learn a better representation of the underlying scale-free nature of social networks (Krioukov et al., 2010; Papadopoulos et al., 2012).

Initialization: Our proposed HGCN aggregates features from neighboring nodes based on graph convolutions in the hyperbolic space to enrich learned language and historical emotion features. We initialize user nodes with their historical emotional spectrum $\mathbf{E^i_j}$ obtained through the HEAT mechanism, and tweet nodes with their emotional representation $\mathbf{T'_i}$. Hyper-SOS then performs hyperbolic graph convolutions on these user and tweet features on the social graph $\mathcal{G}$ with $|U|$ user nodes and $|T|$ tweet nodes, which can also be represented by: its adjacency matrix $\mathbf{A} \in \mathbb{R}^{(|U|+|T|)\times(|U|+|T|)}$, a diagonal degree matrix $\mathbf{D}$, where $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$ and a feature matrix $\mathbf{F}^E \in \mathbb{R}^{(|U|+|T|)\times 768}$ in the Euclidean space (denoted by ^{E}), containing the 768-dimensional representation of each node.

Feature Aggregation by Hyperbolic Graph Convolutions: To capture the network's hierarchical structure, Hyper-SOS first uses Poincaré ball manifold ($\exp_o^K$) with a sectional curvature $-1/K$, to map the features $\mathbf{F}^E$ to the hyperbolic space (denoted by ^{H}) $\mathbf{F}^H = \exp_o^K(\mathbf{F}^E)$ as shown in Figure 3. Next, we perform a linear transformation to capture macroscopic neighborhood structures on the Poincaré ball manifold, followed by a Frechet Mean operation (Fréchet, 1948) denoted by $\texttt{FM}$. Owing to the trainable curvature K, Hyper-SOS utilizes a hyperbolic non-linear activation with varying curvature ($\sigma^{\otimes^{K_{i-1},K_i}}$) to allow a different curvature at each HGCN layer. $\otimes$ is the Möbius transformation operator. Formally, the feature aggregation-based update rule at the i^{th} HGCN layer is:

$$\mathbf{O}^{(i)} = \sigma^{\otimes^{K_{i-1},K_i}}(\texttt{FM}(\tilde{\mathbf{A}}\mathbf{O}^{(i-1)}\mathbf{W}^{(i)})) \quad (2)$$

where $-1/K_{i-1}$ and $-1/K_i$ are the hyperbolic curvatures at layer $i-1$ and i, respectively. $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ is the degree normalized adjacency matrix and $\mathbf{W}$ is a trainable network parameter.

Finally, Hyper-SOS applies a dense layer with Rectified Linear Unit ($ReLU$) to get a prediction vector, followed by softmax to output the probabilities for the presence of SI ($\hat{\mathbf{y}}$) as:

$$\hat{\mathbf{y}} = softmax(ReLU(\mathbf{W}_y(\mathbf{O}^{(2)}) + \mathbf{b}_y)) \quad (3)$$

where, $\{\mathbf{W}_y, \mathbf{b}_y\}$ are network parameters. $\mathbf{O}^{(2)}$ is the output of two stacked convolutions (Equation 2), with input $\mathbf{O}^{(0)}$ set as the initial features F^H.

3.6 Hyper-SOS Training and Optimization

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 HGCN 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:

$$\mathcal{L} = \text{CB}_{focal}(\hat{\mathbf{y}}_\mathbf{i}, y_i; \beta_{cb}, \gamma) \quad (4)$$

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_{cb} and γ are hyperparameters.

4 Dataset Properties

4.1 Data description

We use an existing Twitter dataset curated by Mishra et al. (2019b). The dataset contains Twitter timelines of 32,558 unique users, spanning over ten years of historical tweets from 2009 to 2019, summing up to 2.3M unlabeled tweets. The users were selected based on a seed lexicon of 143 suicidal phrases (e.g., “wanting to die", “last day"), which identified 34,306 tweets potentially containing suicide ideation. Two psychology students then annotated these tweets under the supervision of a professional psychologist, achieving Cohen's κ of 0.72, under the below guidelines:

SI Present: Tweets where suicide ideation or attempts are discussed in a somber, non-flippant tone.
SI Absent: Tweets with no evidence for risk of suicide, e.g., song lyrics, condolences, news.

3984 of the annotated tweets were identified as truly containing suicidal ideation. We feed all the 2.3M tweets to the HEAT mechanism to build user representations (§3.3). The number of historical tweets per user (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.

4.2 Data Split

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

4.3 Network Analysis

In Table 1, we outline quantitative analyses of the social network $\mathcal{G}$ and report Gromov's δ-hyperbolicity of the graph (Jonckheere et al., 2008). A lower hyperbolicity δ indicates a scale-free graph, for trees $\delta = 0$. Based on the low hyperbolicity (Chami et al., 2019), values of the power law coefficients x_{min}, α (Clauset et al., 2009) of the graph $\mathcal{G}$, and the frequency distribution of node degrees in Figure 4, we note that the social network graph $\mathcal{G}$ shows scale-free characteristics. These observations validate our experimental design, and are in line with social network analysis on the structure of social media (Gonçalves et al., 2011), particularly Twitter (Bakshy et al., 2011).

5 Experimental Settings

5.1 Baselines

We reimplement and compare the following previous works to Hyper-SOS on temporal split (§4.2):

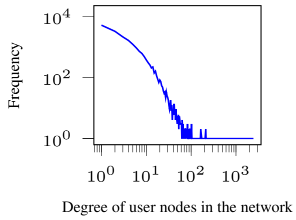


Figure 4: Node degree frequency distribution

Property	Value
Hyperbolicity δ	1.5
Max. Node Degree	2,452
Median Node Degree	1.0
Node Density	$1.9e^{-4}$
Power Law	$p(x) = Cx^{-\alpha}$
x_{min}	14.0
α	2.97

Table 1: Network analysis and statistics

RF + TF (Sawhney et al., 2018): Feeds features such as statistical, LIWC (Pennebaker et al., 2001), n-grams, and POS counts from the tweet to a Random Forest (RF) classifier.

LSTM (Coppersmith et al., 2018): A deep neural network model that uses an LSTM for sequentially encoding GloVe embedding of tweets.

C-CNN (Gaur et al., 2019): Utilizes GloVe encoded tweets as a bag of tweets that are then concatenated and fed non-sequentially to a Contextual Convolutional Neural Network (Shin et al., 2018).

Suicide Detection Model (SDM) (Cao et al., 2019): Applies LSTM + Attention over fine-tuned FastText embeddings of historical tweets, followed by concatenation with tweet to be assessed.

DualContextBert (Matero et al., 2019): Best performing model at CLPsych 2019 (Zirikly et al., 2019). BERT embeddings of each historical tweet are sequentially fed to an attention-based RNN.

STATENet (Sawhney et al., 2020): A deep neural network model. Uses T-LSTM (Baytas et al., 2017) which applies a monotonically decreasing function of elapsed time to weight historical tweets and utilizes BERT fine-tuned on Plutchik-based emotions for the tweet to be assessed.

SNAP-BATNET (Mishra et al., 2019b): Encodes social graph structure using Node2Vec (Grover and Leskovec, 2016) embeddings concatenated with GloVe embeddings for the tweet to be assessed. They report weighted F1.

5.2 Experimental Setup

We evaluate Hyper-SOS using macro F1 score and recall for the SI class. We set hyperparameters for all models based on the validation macro F1 score. We use Grid search to explore: Hidden dimension $H^d \in \{128, 256, \cdots, 1024\}$, Dropout $\delta \in \{0.0, 0.1, \cdots, 0.7\}$. For the HEAT: $\beta \in \{1e^{-3}, \cdots 1e^{-1}\}$ and $\epsilon \in \{1e^{-3}, \cdots 1e^{-1}\}$. $\beta_{cb} \in \{0.999, 0.9999, \cdots, 0.999999\}$ and $\gamma \in \{2.0, 2.5, \cdots, 4.0\}$, learning rate $I_{lr} \in \{1e^{-6}, \cdots, 1e^{-3}\}$, weight decay $w_d \in \{1e^{-6}, \cdots, 1e^{-3}\}$. We find the optimal hyperparameters as: $H_d = 512$, $\delta = 0.2$, $\beta = 1e^{-3}$, $\epsilon = 1e^{-2}$, β_{cb} =0.9999, $\gamma = 3.0$, $I_{lr} = 1e^{-4}$, $w_d = 5e^{-4}$. We use PyTorch for all models, optimize Hyper-SOS using Adam for 5,000 epochs and apply early stopping with a patience of 100 epochs in 1,260s on a Tesla K80 GPU.

6 Results and Analysis

6.1 Comparisons with Prior Work

Type of Context	Model	M. F1 ↑	Recall$_s$ ↑
Non-Contextual	RF+TF	0.536	0.513
	CLSTM	0.588	0.597
Historical Context	CCNN	0.729	0.587
	SDM	0.743	0.755†
	DualContextBERT	0.767	0.786†
	STATENet	0.799*†	0.810*†
Social Context	SNAPBATNET	0.776*	0.606
Social + Historical	HyperSOS	0.792*†	0.818*†

Table 2: Mean of results obtained over 10 runs. * indicates that the result is significantly ($p < 0.005$) better than DualContextBert and † represents better than SNAPBATNET under Wilcoxon's Signed Rank test). **Bold** indicates best performance.

Contextual vs. Non-Contextual Models: We compare Hyper-SOS with a variety of models in Table 2. We categorize the models as *non-contextual*, i.e., using the current tweet only, and more recent *user-contextual*, spanning both *social* and *historical* context. We note that user-contextual models drastically outperform RF+TF and LSTM that only leverage the language of the tweet without any additional user context. We attribute these improvements to the ability of personally contextual models to better ascertain a user's mental state through their historical activity and communities they interact with (Flek, 2020).

Contextual Models: Amongst models utilizing user's historical tweeting activity, we note methods modeling user tweets as temporal sequence (DualContextBERT, Hyper-SOS) outperform bag-of-tweets based models (C-CNN, SDM). On the other hand, prior work leveraging shallow features from social graph's structure without any temporal context (SNAPBATNET), is competitive to historical context models. This sets the premise for leveraging user's social relations as shallow features in neural methods, validating the effectiveness of social context for suicide ideation detection. Hyper-SOS significantly ($p < 0.005$) outperforms both social and historical contextual models, by virtue of its design. Hyper-SOS's design captures the scale-free nature of social relations through deep graph convolutions that blend language features across a user's historical tweeting activity to ascertain suicide ideation. These results validate the potential of utilizing social and historical context, as reflected in psychological works discussing the interpersonal theory of suicide (Joiner, 2007, 2009; Orden et al., 2010). The higher Macro F1 of STATENet can be attributed to its compute-intensive, learnable historical modeling component. We leave using a learnable model to encode personal historical context to our future research directions.

Hyper-SOS advances prior work on multiple fronts: i) combining social and historical context, ii) deep graph convolutions rather than shallow structural features, iii) capturing the scale-free nature of social networks through hyperbolic transformation, and iv) modeling a user's emotions based on the HEAT Mechanism. We explore the impact of each of these design choices through a series of ablative and exploratory analyses next.

6.2 Hyper-SOS Ablation Study

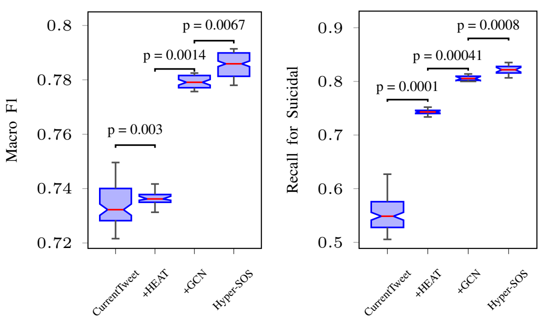


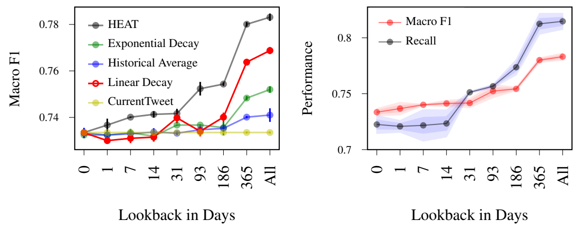
Figure 5: Confidence intervals for evaluation metrics of Ablation study over 10 different runs. (p) indicates the p-value under Wilcoxon's Signed Rank test.

We analyze Hyper-SOS's components through an ablation study in Figure 5. We start by examining how the predictive power of the base (`CurrentTweet`) model changes when enriched with user's historical emotional context (`HEAT`), then gradually with social context (`GCN`), and finally on adding hyperbolic transformations over graph convolutions (`Hyper-SOS`). We note that incorporating a user's historical emotion spectrum via HEAT in a time-sensitive manner improves performance. Specifically, we note improvement in recall in terms of correctly identifying the presence of suicide ideation, likely due to the contextualization of a user's mental state via temporal context.

We note significant ($p < 0.05$) improvements by leveraging social context, learning representations through feature aggregations within a user's neighborhood. These aggregations enrich the learned representations through the structure and historical emotion-based features of the communities the user interacts with, further amplifying the predictive power by greater contextualization.

Lastly, building on the scale-free nature of social networks (Cox et al., 2012), leveraging feature transformations and graph convolutions in the hyperbolic space brings further improvements, as plain GCNs are unable to generalize over such hierarchical scale-free structures (Fronczak, 2018). Our observations revalidate the utility of Hyper-SOS for suicide ideation detection, specifically the influence of social context, and correctly capturing the network's scale-free traits (Rosenquist et al., 2011).

6.3 Impact of Historical Context Aggregation



(a) Other temporal functions (b) HEAT saturation

Figure 6: F1 changes with (a) other temporal user embeddings and (b) different temporal window (10 runs).

We analyze Hyper-SOS's sensitivity to the choice of temporal kernels for aggregating user's historical tweets as shown in Figure 6a. Overall, we notice that all user features learned via temporal aggregations outperform the `CurrentTweet` representation that does not use any historical information. The temporal kernels' performance improves as we factor in more historical tweets up to a year. We also find that Linear Decay performs better than Exponential Decay, hinting towards the importance of older tweets ($>$ 3 months), in some cases, for contextualizing user's more recent suicide ideation with past emotional states.

We note that using the HEAT mechanism as a temporal kernel consistently bestows significant improvements in Hyper-SOS's performance over time compared to all other variants. Self-exciting

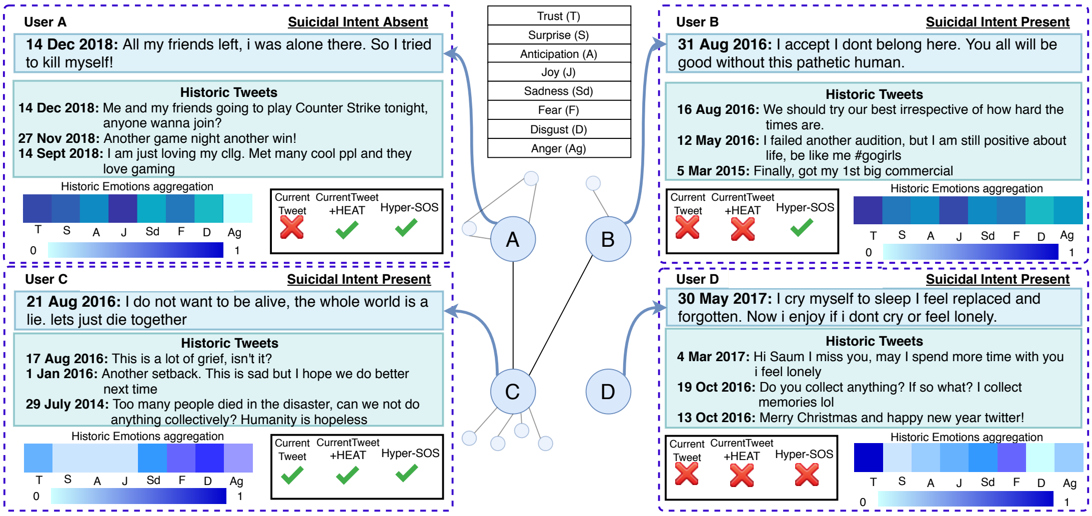


Figure 7: We study four users in a social graph, with their tweet to be assessed, historical tweets, and timestamps. The social graph shows the four users and their interactions among themselves and other users. We also show aggregated historical emotions through HEAT mechanism over time.

temporal point processes such as the Hawkes mechanism have shown great promise in modeling social media dynamics (Rizoiu et al., 2017) and user behavior over time (Guo et al., 2019), revalidating the effectiveness of our proposed HEAT mechanism for learning user representations. Further, we note that Hyper-SOS's performance saturates on adding history beyond a year (Figure 6b). This is in line with psychology research, noting the depreciating importance of user's emotions over longer time periods (Selby et al., 2013; Kaplow et al., 2014; Glenn et al., 2020).

6.4 Impact of Different User Relations

Relation Type	Macro F1 ↑	Recall for SI ↑
All (Hyper-SoS)	**0.792***	**0.818***
(×) Mentions	0.774	0.771
(×) Quotes	0.780	0.804
(×) ReplyTo	0.776	0.802

Table 3: Mean of performance by removing each type of relation from the social network graph obtained over 10 runs. Result with all relations is significantly better than with any relation type removed.

We analyze the importance of different types of social network relations based on how two users interact, by removing each relation type from the graph, as shown in Table 3. We note that removing relations based on user mentions, Hyper-SOS's performance drastically drops. We postulate this drop to the physical and cognitive effort a mention requires, as opposed to other forms of user interactions, in fact, it is the strongest form of user interaction on Twitter (Fink et al., 2016). This observation aligns with the findings of prior social network research that explore the influence of different communications on Twitter (Grabowicz et al., 2012), especially in networks where users can be influenced by a few "known" users (Cha et al., 2010). We note that relatively weaker forms of interactions such as quotes and replies do not contribute towards social context as much as mentions for suicide ideation detection. As suggested in past studies (Sultana et al., 2017), we observe that combining all the user interactions significantly ($p < 0.005$) improves Hyper-SOS's performance.

6.5 Exploratory and Error Analysis

We now present a qualitative analysis (Figure 7) to derive deeper insights into Hyper-SOS's predictive power. We see that the most recent tweet by user A shows explicit signs of suicidal intent. However, from their historical tweets, we notice that User A is talking about their gaming experience. Hence, studying the tweet to be assessed in isolation is not sufficient to assess users' risk, even for humans. Indeed, only temporally contextual models (HEAT, Hyper-SOS) correctly predict the absence of suicidal intent. In a more challenging case, that of user B, the tweet to be assessed shows no overt signs of suicidal intent, and their historical activity is not concerning either. Hyper-SOS's graph-based learning alleviates this issue by learning from a user's social context. Upon analyzing the network, we note User B's interaction with user C's tweets, which are suicidal, which might influence the tendency of User B to show suicidal behavior. Moreover, user C is a highly connected, influential node and has the potential to impression the emotions of users who interact with it (Chung and Zeng, 2020). User D presents an error case. We find that the tweet to be assessed is ambiguous, and historical activity is not informative either. Moreover, user D is isolated, highlighting that suicide ideation detection in the absence of contextual elements (historical activity, network interactions) can be highly subjective, and paves the way for future work.

7 Broader Impact and Ethics

Emphasizing the sensitive nature of this work, we acknowledge the trade-off between privacy and effectiveness (Eskisabel-Azpiazu et al., 2017). To avoid coercion and intrusive treatment, we work within the purview of acceptable privacy practices suggested by Chancellor et al. (2019b) and considerations discussed by Fiesler and Proferes (2018). Although informed consent of each user was not sought as it may be deemed coercive, we perform automatic de-identification of the dataset using named entity recognition (Benton et al., 2017a,b) to reduce the risk of including any identifying data in the raw data. We paraphrase all examples shown in this work to protect user privacy (Chancellor et al., 2019a,b). All the user data is kept separately on protected servers linked to the raw text and network data only through anonymous IDs.

We acknowledge that it is almost impossible to prevent abuse of released technology even when developed with good intentions (Jonas, 1984; Hovy and Spruit, 2016). Hence, we ensure that this analysis is shared only selectively and subject to IRB approval (Zimmer, 2009, 2010) to avoid misuse such as Samaritan's Radar (Hsin et al., 2016).

Limitations: We acknowledge that suicidality is subjective (Keilp et al., 2012), the interpretation of this analysis may vary across individuals on social media (Puschman, 2017), and we do not know the true intentions of the user behind the post. We further acknowledge that suicide risk exists on a diverse spectrum (Bryan and Rudd, 2006), and a binary distinction is a task simplification intended to alert the human in the loop about exceeding a possible intervention threshold. We also recognize that the studied data may be susceptible to demographic, annotator, and medium-specific biases (Hovy and Spruit, 2016).

Future Practical Applicability In the future, we would want to focus on creating a differentially private public model that can be shared with the community while preserving user privacy (Lyu et al., 2020; Yu et al., 2019). Further, suicide ideation detection on social media can involve failure modes that could potentially incorrectly ascertain suicide risk. To this end, we focus on Hyper-SOS as a preliminary tool for prioritizing human expert, clinical psychologist-based assessment.

8 Conclusion

Motivated by psychological studies, we propose a framework jointly leveraging emotional history from user's past tweets and social information from user's neighborhood in a network to contextualize the interpretation of the latest tweet of a user. To our knowledge, this is the first deep graph neural network study to automatically identify suicide ideation on social media. Reflecting upon the scale-free nature of social network relationships, we propose the use of Hyperbolic Graph Convolution Networks, and demonstrate that these are more suitable for our Twitter task than their euclidean counterparts. Inspired by geophysics, we further propose the use of HEAT Mechanism to learn the historic emotional spectrum of a user in a time-sensitive manner. When analyzing the contributions of its individual components to assess suicidal intent, we demonstrate the beneficial impact of both the social and personal context representations.

Acknowledgements

This work has been supported by the German Federal Ministry of Education and Research (BMBF) as a part of the Junior AI Scientists program under the reference 01-S20060. We thank the anonymous reviewers for their valuable input.

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.

Sofía Aparicio, Javier Villazón-Terrazas, and Gonzalo Álvarez. 2015. A model for scale-free networks: application to twitter. *Entropy*, 17(8):5848–5867.

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.

Chen Avin, Zvi Lotker, David Peleg, Yvonne-Anne Pignolet, and Itzik Turkel. 2018. Elites in social networks: An axiomatic approach to power balance and price's square root law. *PloS one*, 13(10):e0205820.

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

JinYeong Bak, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 60–64.

Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Identifying influencers on twitter. In *Fourth ACM International Conference on Web Seach and Data Mining (WSDM)*.

Albert-László Barabási and Eric Bonabeau. 2003. Scale-free networks. *Scientific american*, 288(5):60–69.

Philip J Batterham, Alison L Calear, and Helen Christensen. 2013. Correlates of suicide stigma and suicide literacy in the community. *Suicide and Life-Threatening Behavior*, 43(4):406–417.

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.

David R Bild, Yue Liu, Robert P Dick, Z Morley Mao, and Dan S Wallach. 2015. Aggregate characterization of user behavior in twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology (TOIT)*, 15(1):1–24.

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.

Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, page 75–84, New York, NY, USA. Association for Computing Machinery.

John T Cacioppo, James H Fowler, and Nicholas A Christakis. 2009. Alone in the crowd: the structure and spread of loneliness in a large social network. *Journal of personality and social psychology*, 97(6):977.

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.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, P Krishna Gummadi, et al. 2010. Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. 2019. Hyperbolic graph convolutional neural networks. In *Advances in neural information processing systems*, pages 4868–4879.

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.

Wei Chen, Wenjie Fang, Guangda Hu, and Michael W Mahoney. 2013. On the hyperbolicity of small-world and treelike random graphs. *Internet Mathematics*, 9(4):434–491.

Wingyan Chung and Daniel Zeng. 2020. Dissecting emotion and user influence in social media communities: An interaction modeling approach. *Information & Management*, 57(1):103108.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Gualtiero B Colombo, Pete Burnap, Andrei Hodorog, and Jonathan Scourfield. 2016. Analysing the connectivity and communication of suicidal users on twitter. *Computer communications*, 73:291–300.

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.

Georgina R Cox, Jo Robinson, Michelle Williamson, Anne Lockley, Yee Tak Derek Cheung, and Jane Pirkis. 2012. Suicide clusters in young people: evidence for the effectiveness of postvention strategies. *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, 33(4):208.

Arthur H Crisp, Michael G Gelder, Susannah Rix, Howard I Meltzer, and Olwen J Rowlands. 2000. Stigmatisation of people with mental illnesses. *The British journal of psychiatry*, 177(1):4–7.

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.

Ewa K Czyz, Adam G Horwitz, Daniel Eisenberg, Anne Kramer, and Cheryl A King. 2013. Self-reported barriers to professional help seeking among college students at elevated risk for suicide. *Journal of American college health*, 61(7):398–406.

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.

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in NLP. 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 4707–4717, Hong Kong, China. Association for Computational Linguistics.

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.

Robert A Fahey, Tetsuya Matsubayashi, and Michiko Ueda. 2018. Tracking the werther effect on social media: Emotional responses to prominent suicide deaths on twitter and subsequent increases in suicide. *Social Science & Medicine*, 219:19–29.

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

Clay Fink, Aurora Schmidt, Vladimir Barash, Christopher Cameron, and Michael Macy. 2016. Complex contagions and the diffusion of popular twitter hashtags in nigeria. *Social Network Analysis and Mining*, 6(1):1.

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.

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.

Maurice René Fréchet. 1948. Random elements of any kind in a remote space. *Annals of the Henri Poincaré institute*, 10(4):215–310.

Piotr Fronczak. 2018. *Scale-Free Nature of Social Networks*, pages 2300–2309. Springer New York, New York, NY.

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.

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.

Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. 2011. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, 6(8):e22656.

Przemyslaw A Grabowicz, José J Ramasco, Esteban Moro, Josep M Pujol, and Victor M Eguiluz. 2012. Social features of online networks: The strength of intermediary ties in online social media. *PloS one*, 7(1):e29358.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Siwen Guo, Sviatlana Höhn, and Christoph Schommer. 2019. A personalized sentiment model with textual and contextual information. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 992–1001, Hong Kong, China. Association for Computational Linguistics.

Yari Gvion and Alan Apter. 2012. Suicide and suicidal behavior. *Public health reviews*, 34(2):9.

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

Keith M Harris and Melissa Ting-Ting Goh. 2017. Is suicide assessment harmful to participants? findings from a randomized controlled trial. *International journal of mental health nursing*, 26(2):181–190.

Alan G Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

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

Joseph Henrich and Francisco J Gil-White. 2001. The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and human behavior*, 22(3):165–196.

Alison L Hill, David G Rand, Martin A Nowak, and Nicholas A Christakis. 2010. Emotions as infectious diseases in a large social network: the sisa model. *Proceedings of the Royal Society B: Biological Sciences*, 277(1701):3827–3835.

Elizabeth M Hill, Frances E Griffiths, and Thomas House. 2015. Spreading of healthy mood in adolescent social networks. *Proceedings of the Royal Society B: Biological Sciences*, 282(1813):20151180.

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.

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, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*.

Thomas Joiner. 2007. *Why people die by suicide*. Harvard University Press.

Thomas Joiner. 2009. Psychological science agenda, june 2009. *Psychological Science*.

Hans Jonas. 1984. The imperative of responsibility: In search of an ethics for the technological age.

Klaus Jonas. 1992. Modelling and suicide: a test of the werther effect. *British Journal of Social Psychology*, 31(4):295–306.

Edmond Jonckheere, Poonsuk Lohsoonthorn, and Francis Bonahon. 2008. Scaled gromov hyperbolic graphs. *Journal of Graph Theory*, 57(2):157–180.

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.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. 2010. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106.

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

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.

David D Luxton, Jennifer D June, and Jonathan M Fairall. 2012. Social media and suicide: a public health perspective. *American journal of public health*, 102(S2):S195–S200.

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020. Differentially private representation for NLP: Formal guarantee and an empirical study on privacy and fairness. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2355–2365, Online. Association for Computational Linguistics.

J John Mann. 2002. A current perspective of suicide and attempted suicide. *Annals of internal medicine*, 136(4):302–311.

Naoki Masuda, Issei Kurahashi, and Hiroko Onari. 2013. Suicide ideation of individuals in online social networks. *PloS one*, 8(4):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.

Michael J McCarthy. 2010. Internet monitoring of suicide risk in the population. *Journal of affective disorders*, 122(3):277–279.

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).

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Alex Mesoudi. 2009. The cultural dynamics of copycat suicide. *PLoS One*, 4(9):e7252.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019a. Abusive language detection with graph convolutional networks. 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 2145–2150.

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019b. 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.

Anna S Mueller and Seth Abrutyn. 2015. Suicidal disclosures among friends: using social network data to understand suicide contagion. *Journal of health and social behavior*, 56(1):131–148.

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.

Kimberly A Van Orden, Tracy K Witte, Kelly C Cukrowicz, Scott R Braithwaite, Edward A Selby, and Thomas E Joiner Jr. 2010. The interpersonal theory of suicide. *Psychological review*, 117(2):575.

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.

Fragkiskos Papadopoulos, Maksim Kitsak, M Ángeles Serrano, Marián Boguná, and Dmitri Krioukov. 2012. Popularity versus similarity in growing networks. *Nature*, 489(7417):537–540.

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.

John P Pestian, Michael Sorter, Brian Connolly, Kevin Bretonnel Cohen, Cheryl McCullumsmith, Jeffry T Gee, Louis-Philippe Morency, Stefan Scherer, Lesley Rohlfs, and STM Research Group. 2017. 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.

Erzsébet Ravasz and Albert-László Barabási. 2003. Hierarchical organization in complex networks. *Physical review E*, 67(2):026112.

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

Marian-Andrei Rizoiu, Lexing Xie, Scott Sanner, Manuel Cebrian, Honglin Yu, and Pascal Van Hentenryck. 2017. Expecting to be hip: Hawkes intensity processes for social media popularity. In *Proceedings of the 26th International Conference on World Wide Web*, pages 735–744.

Lindsay Robertson, Keren Skegg, Marion Poore, Sheila Williams, and Barry Taylor. 2012. An adolescent suicide cluster and the possible role of electronic communication technology. *Crisis*.

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.

J Niels Rosenquist, James H Fowler, and Nicholas A Christakis. 2011. Social network determinants of depression. *Molecular psychiatry*, 16(3):273–281.

James N Rosenquist. 2011. Lessons from social network analyses for behavioral medicine. *Current Opinion in Psychiatry*, 24(2):139–143.

Gregory A Roth, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. 2018. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1736–1788.

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

Marialisa Scatà, Alessandro Di Stefano, Aurelio La Corte, and Pietro Liò. 2018. Quantifying the propagation of distress and mental disorders in social networks. *Scientific reports*, 8(1):1–12.

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.

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.

Madeena Sultana, Padma Polash, and Marina Gavrilova. 2017. Authorship recognition of tweets: A comparison between social behavior and linguistic profiles. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 471–476. IEEE.

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.

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.

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

Matt Wray, Cynthia Colen, and Bernice Pescosolido. 2011. The sociology of suicide. *Annual Review of Sociology*, 37.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

L. Yu, L. Liu, C. Pu, M. E. Gursoy, and S. Truex. 2019. Differentially private model publishing for deep learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 332–349.

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.