RISE: Robust Early-exiting Internal Classifiers for Suicide Risk Evaluation

Ritesh Soun[★], **Atula Neerkaje**[†], **Ramit Sawhney**[◇,▽], **Nikolaos Aletras**[♣], **Preslav Nakov**[▽]

[★]Sri Venkateswara College, [†]The University of Texas at Austin, [◇]Georgia Institute of Technology,
[▽]Mohamed bin Zayed University of Artificial Intelligence, [♣]The University of Sheffield
sounritesh@gmail.com, atutej@utexas.edu, n.aletras@sheffield.ac.uk
{ramit.sawhney, preslav.nakov}@mbzuai.ac.ae

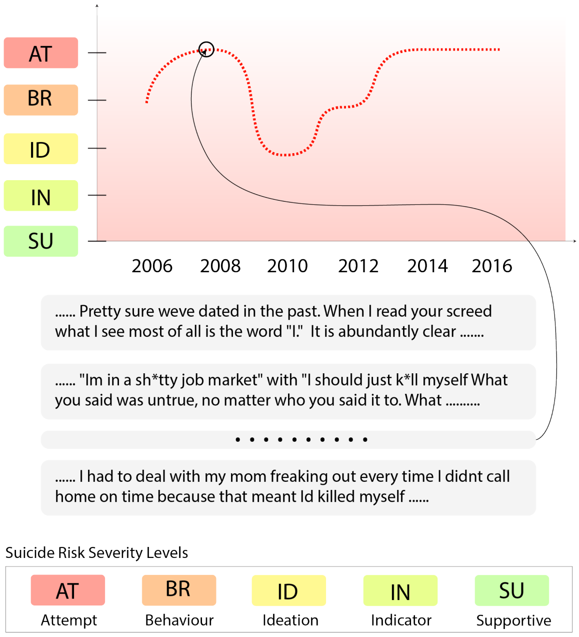
Abstract

Suicide is a serious public health issue, but it is preventable with timely intervention. Emerging studies have suggested there is a noticeable increase in the number of individuals sharing suicidal thoughts online. As a result, utilising advance Natural Language Processing techniques to build automated systems for risk assessment is a viable alternative. However, existing systems are prone to incorrectly predicting risk severity and have no early detection mechanisms. Therefore, we propose RISE, a novel robust mechanism for accurate early detection of suicide risk by ensembling Hyperbolic Internal Classifiers equipped with an abstention mechanism and early-exit inference capabilities. Through quantitative, qualitative and ablative experiments, we demonstrate RISE as an efficient and robust human-in-the-loop approach for risk assessment over the Columbia Suicide Severity Risk Scale (C-SSRS) and CLPsych 2022 datasets. It is able to successfully abstain from 84% incorrect predictions on Reddit data while out-predicting state of the art models upto 3.5x earlier.

Keywords: Suicide Risk Evaluation, Internal Classifiers, Hyperbolic Learning

1. Introduction

Upwards of 700,000 people die due to suicide every year (WHO, 2021). It is the fourth leading cause of death among 15-29 year olds and is the second leading cause of death among 10-14 year olds in America (CDC, 2021). Suicide is a serious public health problem but it is preventable with timely intervention and professional help. However, over two-thirds of individuals who die from suicide do not seek professional mental health support (Stene-Larsen and Reneflot, 2019). 39% of those who do seek professional help, do not disclose suicidal intent to their therapists (McGillivray et al., 2022).



...... Pretty sure weve dated in the past. When I read your screed what I see most of all is the word "I." It is abundantly clear ....

...... "Im in a sh*tty job market" with "I should just k*ll myself What you said was untrue, no matter who you said it to. What ....

• • • • • • • • • •

...... I had to deal with my mom freaking out every time I didnt call home on time because that meant Id killed myself ......

Figure 1: We visualize the suicide risk severity of a sample of Reddit posts for a user from the CSSRS Dataset with "Attempt (AT)" risk severity and plot it over time.

Previous studies have shown that with widespread use of social media (Chaffey, 2023), people with suicidal intent may disclose suicidal thoughts or seek information for support online (Fahey et al., 2020; Daine et al., 2013; Colombo et al., 2016). In addition, studies also suggest a notable increase in the occurrence of young individuals expressing suicidal thoughts by posting suicidal notes on social media platforms (Desmet and Hoste, 2013; Ji et al., 2020). Therefore, online expression of suicidal thoughts is an information rich source for timely detection of suicide risk and is known to be associated with psychologically assessed suicide risk (Coppersmith et al., 2018; Jashinsky et al., 2014). As a result, there has been a growing interest in using Artificial Intelligence (AI) based systems for mental health to create early interventions for patients with chronic mental health conditions (Roy et al., 2022).

In a safety-critical scenario such as mental health, technological robustness is extremely important (Sawhney et al., 2022b), highlighting the need for safe and responsible AI models for such tasks (Garg, 2023), such as the ability to abstain from

LREC-COLING 2024, pages 14134–14145
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

making a prediction (Sawhney et al., 2022b). Additionally, it is important for an AI model to confidently know as early as possible if a user is at risk of suicide (Leiva and Freire, 2017; Smys and Raj, 2021), or in the worst case - know very early that it is uncertain. Existing studies (Van Dijk, 1977; Sawhney et al., 2021a) also show that in data that comprises of extensive long-form content, such as posts on social media, only a handful of key data points exert the strongest influences on the overall trend, which is effectively captured by hyperbolic networks (Agarwal et al., 2022).

As shown in Figure 1, certain key phrases show strong signs of high suicide risk early on, showing that early detection is of immense significance. Therefore, models equipped with such capabilities would need to be endowed with mechanisms that generate early predictions along the temporal dimension (Hochreiter and Schmidhuber, 1997), and with powerful representational power to capture the rare, scale-free (Broido and Clauset, 2019) excitations induced by important texts. Motivated by psychological studies, and building on prior work, we summarize our contributions as:

- We model the hyperbolic nature of online text streams using a better suited geometry that captures the powerlaw dynamics in social media texts (section 4.1).

- We formulate RISE, a novel risk-averse mechanism for early detection of suicide risk by ensembling Hyperbolic Internal Classifiers (section 4.3) equipped with an abstention mechanism (section 4.4) and early-exit inference capabilities (section 4.6).

- Through ablative (section 6.2), qualitative (section 6.4) and quantitative (section 6.1, section 6.3) experiments, we demonstrate the ability of RISE as a robust and efficient approach for early detection of suicide risk using online text streams.

2. Related Work

Text stream modelling helps in detecting patterns from a sequence of textual data such as social media posts. Analysing such sequences in succession provides better contextual representation (Hu et al., 2018) due to the sequential context dependency present in them. While, this has proven to be effective in the healthcare domain in the past (Lampos et al., 2010; Paul and Dredze, 2021; Baytas et al., 2017), it comes with its own challenges. Social theories indicate that only a few texts have a substantial influence on the overall trend, following a power-law distribution (Van Dijk, 1977; Gabaix, 2016). These influential texts are rare and exhibit scale-free properties (Zhao et al., 2010). Modeling such power-law dynamics is challenging due to their hierarchical nature (Sala et al., 2018). Hyperbolic learning has shown promise in capturing power-law dynamics in various domains, including computer vision (Khrulkov et al., 2020) and natural language processing (Tifrea et al., 2019). Recent work (Agarwal et al., 2022) leverage the power-law dynamics of text streams (Gabaix, 2016; Van Dijk, 1977) and their varying impact on different events through hyperbolic learning (Sala et al., 2018). When applied to applications in mental health (Sawhney et al., 2022a; Agarwal et al., 2022), hyperbolic learning has significantly advanced state-of-the-art. In recent years, Natural Language Processing (NLP) has shown great promise for suicide risk assessment based on online user behavior (Sawhney et al., 2022a; Ji et al., 2020; De Choudhury et al., 2016; Coppersmith et al., 2014). Such approaches have proven useful for the social NLP research community to analyse and understand associations between users' social media posts and mental health status (Garg, 2023). A drawback of these methods is that they are inherently designed to predict even when uncertain, posing a risk for mental health applications which are safety critical.

Recently, (Sawhney et al., 2022b) explored an approach for suicide risk assessment from the perspective of selective classification (Ziyin et al., 2020), where a model was trained to abstain from making predictions when not certain. While this enables a human in the loop to interpret the low confidence level of the model and intervene if deemed necessary, it does not allow the model to generate early predictions when the confidence is high to plan interventions in advance. Emerging studies with early exit mechanisms in pre-trained language models (Sun et al., 2021; Liao et al., 2021; Xin et al., 2020) have demonstrated efficiency gains by introducing internal classifiers at each layer allowing the model to completely rely on the first few hidden layers of the model to make predictions whenever possible (Sun et al., 2022; Zhou et al., 2020). While these studies explore early exiting in language models, they do not attempt to apply it to tasks that have a temporal dimension and scale-free properties. Time-critical tasks like suicide risk assessment (Leiva and Freire, 2017; Smys and Raj, 2021) can utilise an early-exiting mechanism to allow the model to make early predictions and exit early along the temporal dimension, utilising a relatively smaller portion of the text sequence (lesser time-steps) in the process.

Risk level	% Samples
Supportive	20
Indicator	20
Ideation	34
Behaviour	15
Attempt	9

(a) CSSRS Dataset

Risk level	% Samples
Low	22
Moderate	37
High	41

(b) CLPsych 2022 Dataset

Table 1: Percentage distribution of user samples based on risk levels.

3. Datasets and Task

3.1. Datasets

Columbia Suicide Severity Risk Dataset: The Columbia Suicide Severity Risk Scale (C-SSRS) is a widely used questionnaire utilized by psychiatrists to assess the severity of suicide risk (Posner et al., 2011). Unlike in a clinical setting, on social media non-suicidal users may also participate in discussions to offer support to others who are deemed suicidal (Gaur et al., 2021). To address these challenges, additional classes have been defined in the C-SSRS scale (Posner et al., 2011). These include Suicide Indicator and Supportive (Negative class). As released by (Gaur et al., 2019), this dataset comprises Reddit posts from 500 users filtered from an initial set of $270,000$ users across various suicide-related subreddits. The users were annotated by practicing psychiatrists into five risk levels based on the Columbia Suicide Severity Risk Scale (Posner et al., 2011). The average pairwise agreement among the annotators was found to be 0.79, with a group-wise agreement of 0.73, indicating acceptable inter-rater reliability.

CLPsych 2022 Dataset, released by (Tsakalidis et al., 2022) comprises of $6,195$ posts by 185 users from mental health related subreddits (MHS). These 185 users were filtered from an initial set of $83,000$ users having at least 10 posts on MHS.The users were classified into four risk severities - no, low, moderate and high risk by clinical psychology experts. The no and low risk classes were further clubbed into one class due to very low number of samples in the no risk class.

The posts in both these datasets are predominantly in English and the distribution of users among the risk levels for both datasets is given in table 1.

3.2. Task

Following (Gaur et al., 2019; Tsakalidis et al., 2022), we define the task as a multi-class classification problem to predict Y, referring to the suicidal risk of the user $u_i \in \{u_1, u_2, ..., u_N\}$ in increasing order of severity risk whose posts $P_i = \{p^i_1, p^i_2, ..., p^i_T\}$ are in chronological order, with the latest post being p^i_T. Our aim is to expand the label space to $Y \cup \{$Abstain (AB)$\}$ to allow the model to refrain from making a prediction when uncertain.

As a result, $Y \in \{$Support (SU), Indicator (IN), Ideation (ID), Behaviour (BR), Attempt (AT), Abstain (AB)$\}$ for the CSSRS Dataset (Gaur et al., 2019) and $Y \in \{$Low (L), Moderate (M), High (H), Abstain (AB)$\}$ for the CLPsych 2022 dataset (Tsakalidis et al., 2022).

4. Methodology

4.1. Hyperbolic Geometry

The text sequences found in social media datasets often exhibit tree-like hierarchical structures (Sawhney et al., 2021b). As a result, the utilization of hyperbolic geometry effectively capture the intrinsic properties of such data (Sala et al., 2018).

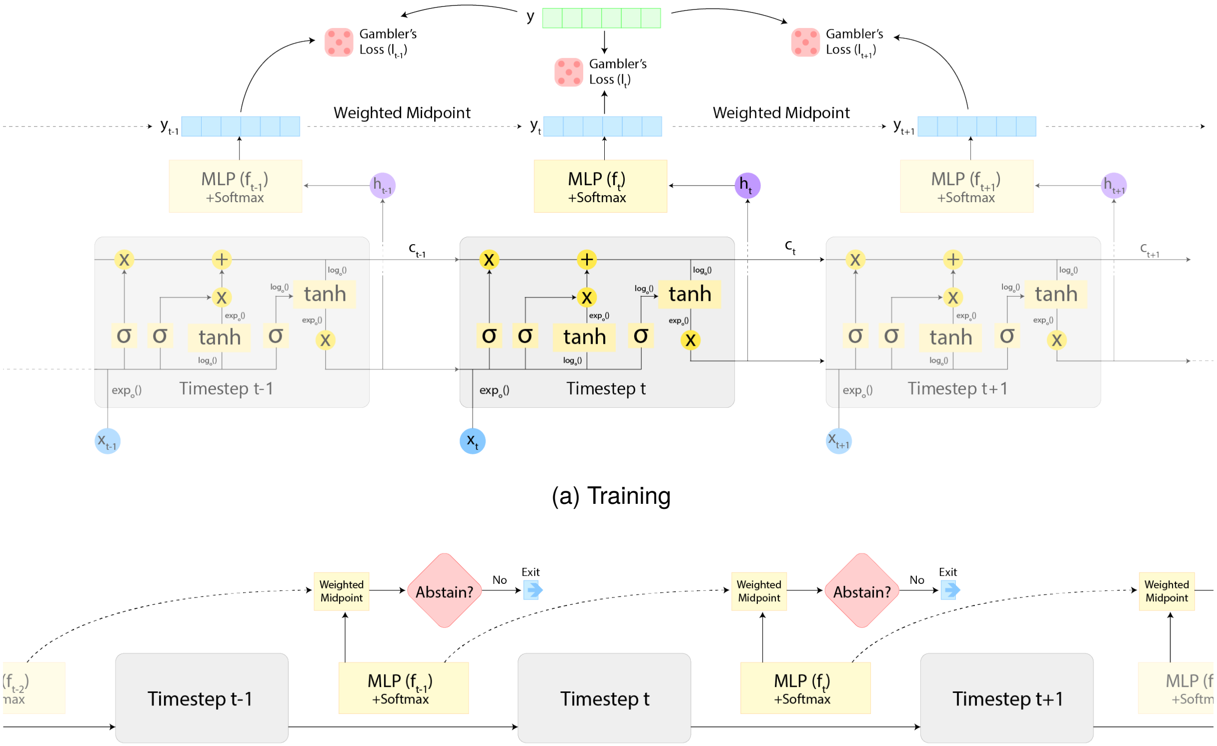
Indeed, the volume of hyperbolic geometry grows exponentially, in contrast to Euclidean spaces where the growth is polynomial (Khrulkov et al., 2020), enabling hyperbolic spaces to capture the underlying scale-free properties of streams (Sala et al., 2018). However, text sequences exhibit a varying degree of scale-free dynamics, which a single geometry cannot capture (Gu et al., 2019). Thus, we seek to learn the optimal underlying geometry. The hyperbolic space is a non-Euclidean space with a constant negative curvature c. To learn the optimal geometry, we aim to learn the curvature c, which controls the degree of hyperbolic properties represented by the space (Gu et al., 2019). We define hyperbolic geometry following (Ganea et al., 2018) and generalize Euclidean operations to the hyperbolic space via Möbius operations given in Ganea et al. (2018).

4.2. Text Embedding Layer

We use Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) to encode each text p^i_k referring to k_{th} post of the i^{th} user to features, $\hat{\boldsymbol{m}}^{\boldsymbol{i}}_{\boldsymbol{k}} =$ BERT$(p^i_k) \in \mathbb{R}^d$ where $d = 768$, obtained by averaging the token level outputs from the final layer of BERT. To apply hyperbolic operations over text features $\hat{\boldsymbol{m}}^{\boldsymbol{i}}_{\boldsymbol{k}}$, we project it to the hyperbolic space via the exponential mapping $\exp_o(\cdot)$ given by, $\boldsymbol{m}^{\boldsymbol{i}}_{\boldsymbol{k}} = \exp_o(\hat{\boldsymbol{m}}^{\boldsymbol{i}}_{\boldsymbol{k}})$

4.3. Hyperbolic LSTM with Internal Classifiers

Hyperbolic LSTMs have been shown to be effective in modeling hierarchical structures and capturing long-range dependencies in sequential data (Nickel and Kiela, 2017; Ganea et al., 2018). We propose a



(a) Training

(b) Inference

Figure 2: An overview of the training and inference mechanism of RISE comprising of an abstention mechanism using Gambler's Loss, Hyperbolic Internal Classifiers and an Early exit inference mechanism

novel variant called Hyperbolic LSTM with Internal Classifiers (HLSTM-IC) that combines the effectiveness of hyperbolic geometry with a classification head at every time step for early detection of suicidal intent.

HLSTM-IC extends the traditional LSTM architecture by incorporating hyperbolic representations into the hidden state update and gating mechanisms, similar to previous hyperbolic LSTM approaches (Nickel and Kiela, 2017; Ganea et al., 2018). The hidden state update and gating functions are modified to operate in the hyperbolic space using the Poincaré disk model, which allows HLSTM-IC to capture hierarchical and tree-like structures in social media posts more effectively. Using hyperbolic features m, we define the current hidden state and current memory states of HLSTM-IC as:

$$\begin{aligned}
\widetilde{\boldsymbol{c}}_{\boldsymbol{t}} &= \sigma \log_{\boldsymbol{o}}(\boldsymbol{W}^c \otimes \boldsymbol{h}_{t-1} \oplus \boldsymbol{U}^c \otimes \boldsymbol{m}_t \oplus \boldsymbol{b}^c) \\
\boldsymbol{C}_t &= \boldsymbol{i}_t \odot \widetilde{\boldsymbol{c}}_{\boldsymbol{t}} \oplus \boldsymbol{f}_t \odot \boldsymbol{C}_{t-1} \quad \text{Current memory} \qquad (1) \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \exp_{\boldsymbol{o}}(\tanh(\boldsymbol{C}_t)) \quad \text{Current hidden state}
\end{aligned}$$

In addition to the hyperbolic LSTM cell, HLSTM-IC comprises of a multilayer perceptron for classification at each time step. The output of the hyperbolic LSTM cell at each time step is passed through the MLP, which acts as an internal classifier to make predictions based on the current state of the sequence. The MLP consists of multiple layers of rectified linear units (ReLUs), followed by a softmax activation to obtain the prediction vector $\hat{y}_t$ for timestep t, given as:

$$\begin{aligned}
\hat{y}_t &= f_t(h_t), \text{ where} \\
f_t(h_t) &= \text{Softmax}(\text{MLP}(h_t))
\end{aligned} \qquad (2)$$

The use of internal classification heads in HLSTM-IC allows the model to make predictions at each time step, enabling early exit decisions based on intermediate predictions, potentially leading to early detection of suicidal intent for high-risk users.

4.4. Abstention Mechanism

To formulate a more robust and fail-safe model, we modify the classification heads to make predictions only when they have a high degree of confidence (Liu et al., 2019) by augmenting the label space with an option to abstain.

Classification heads at each time step output $|Y| + 1$ logits where Y refers to one of the 5 risk severities according to C-SSRS. The extra logit s acts as a selection parameter such that model prediction z_t, for time step t is given as,

$$z_t = (f_t, s) := \begin{cases} \text{Abstain, if } s >= \alpha \\ \text{argmax}(\hat{y}_t), \text{ otherwise} \end{cases} \quad (3)$$

where $\alpha \in (0, 1)$ is the selection threshold.

The selection threshold, α is only used during inference,

- to determine which step to exit at, discussed in detail in later.
- to calculate data coverage $\mathcal{C}$, fraction of the sample space on which predictions are made, abstaining from predicting $1 - \mathcal{C}$ samples. These $1 - \mathcal{C}$ samples can then be manually examined by mental healthcare professionals to identify suicidal intent.

4.5. Joint Network Optimization using Ensembling approach

Each internal classification head, (f_t, s) is trained to predict the ground truth. To take full advantage of this fact (Sun et al., 2021), we construct an ensemble of these classifiers instead of training them independently to optimize the network (Sun et al., 2021; Liao et al., 2021).

To compute the output distribution at time step t, we aggregate the output distributions of the first t time steps. We sequentially calculate a weighted sum at every step $\in \{2, 3, ..., t\}$ using the current and previous output distribution, and decaying factor for the time step given as:

$$\gamma_t = \beta * \gamma_{t-1} + (1 - \beta) \quad (4)$$

where $\beta = 0.5$ is a constant.

Thus, the modified output of the model at time step t (Equation 3) is given as:

$$\hat{z}_t = (1 - \gamma_t) * \text{argmax}(\hat{y}_{t-1}) + \gamma_t * \text{argmax}(\hat{y}_t) \quad (5)$$

where $\hat{z}_t$ is the joint output distribution of the first t internal classifiers.

We can perform an $(m + 1)$-class classification for any m-class classification problem and use the $(m+1)^{th}$ class as an abstention score (Geifman and El-Yaniv, 2019, 2017; Liu et al., 2019). Such models are learnt differently to account for the abstention option and hence, we use Gambler's Loss (Liu et al., 2019) to train our model.

Gambler's Loss corresponding to a particular time step t is given as,

$$\mathcal{L}_t = -\sum_{i}^{|Y|} y_t^i * log(\hat{z}_t^i * r + s) \quad (6)$$

where y_t^i is the ground truth for the t^{th} time step and r is a hyperparameter. A higher value of r discourages abstention. This allows the gradients to propagate through s by refraining from assigning weights to any of the m classes. As a result, s is learnt directly using Gambler's Loss and does not require an extra logit during training which makes it independent of coverage $\mathcal{C}$.

To maximise the likelihood of ground truth Y, we train all internal classifiers using loss $\mathcal{L}$, which is the sum of the losses of internal classifiers, given as:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \ ... \ + \mathcal{L}_T \quad (7)$$

4.6. Early Exit Inference Mechanism

Early detection of suicidal intent is crucial (WHO, 2021), especially for high-risk users to provide timely assistance.

Our early-exit inference mechanism employs internal classifiers. Following Equation 3, our model makes a prediction at every time step. If the selection parameter $s < \alpha$ at time step t, the classifier makes a confident prediction and exits at the current time step, concluding the inference process without the need to go through all time steps (posts).

If our model chooses to abstain, the inference process propagates forward to the next time step. If the exit condition is never reached, our model defaults into the common case of inference in which the complete forward propagation takes place (i.e. the model utilises all posts).

At the end of the complete forward pass, if $s > \alpha$, our model abstains from predicting the sample. Such samples belong to the abstention class $(1 - C)$ and can be directly evaluated by mental health professionals on priority.

5. Experimental Setup

5.1. Training Setup

We have performed all our experiments on a Tesla GPU. We performed a grid search for all our models and selected the best values based on the validation loss. We followed the same preprocessing techniques as suggested by the dataset authors (Gaur et al., 2019; Tsakalidis et al., 2022). We explored the timestep threshold $\tau \in [2, 16]$ and the hidden state dimensions in $\in (64, 128, 256)$. We grid searched our learning rates in $\in (1e-5, 5e-4, 1e-3)$. We used Riemannian Adam (Bécigneul and Ganea, 2018) as our optimizer and a train, dev and test split of 75%, 15% and 10% respectively for both datasets.

5.2. Evaluation Metrics

5.2.1. Grade Precision, Recall and F1-Score

We redefine the metrics for evaluating the model's performance following (Gaur et al., 2019) in order to provide a more accurate assessment. False Positive (FP), represents the ratio of instances where the predicted suicide risk severity level (y^p) is higher than the actual level (y^a) over the size of the test data (NT), and False Negative (FN) captures the ratio of instances where y^p is lower than y^a over NT.

$$\begin{aligned}\text{FP} &= \frac{\sum_{i=1}^{N} I(y_i^p > y_i^a)}{N} \\ \text{FN} &= \frac{\sum_{i=1}^{N} I(y_i^p < y_i^a)}{N}\end{aligned} \tag{8}$$

5.2.2. Fail-safe Rejects

Fail-safe Rejects is the fraction of erroneous abstained samples and is given as the ratio of number of incorrect prediction by the number of samples abstained:

$$\text{Fail-safe Rejects} = \frac{P_{\text{in}}}{P_{\text{abstain}}} \tag{9}$$

5.2.3. Robustness

Robustness is quantified as the fraction of samples correctly classified or abstained for direct evaluation by mental health professionals.

$$\text{Robustness} = \frac{P_{\text{corr+abstain}}}{P_T} \tag{10}$$

5.2.4. Early Detection Efficiency Ratio (EDER)

To quantify how early a model is able to correctly classify samples, we modify Speed-up Ratio from (Xin et al., 2020) and formulate EDER. EDER is defined as the ratio of complete required time steps for an N-step model to the actually executed time steps in the forward pass given as:

$$\text{EDER} = \frac{\sum_{t=1}^{N} N * m_t}{\sum_{t=1}^{t} t * m_t} \tag{11}$$

where m_t is the number of samples that exit at the t^{th} time step. A higher EDER corresponds to a model that can predict the correct class with fewer time steps, i.e. fewer posts corresponding to each user.

6. Results

6.1. Performance Comparison

We compare the performance of RISE with other state-of-the-art methods in Table 2 across two datasets described in Section 3. Contextual CNN (Kim, 2014), using a bag-of-posts approach and SDM (Cao et al., 2019) come out as worst performers. Context Bert (Matero et al., 2019), LSTM (Hochreiter and Schmidhuber, 1997) and n-BiLSTM (Zhang and Rao, 2020) show improvements over Contextual CNN and SDM due to their sequential nature with n-BiLSTM being the best amongst the lot, having a 3% and 2% higher F. score than Contextual CNN for the CSSRS and CLPsych datasets respectively. SISMO (Sawhney et al., 2021c) shows a further increment of 1% in F. score for both datasets as it is able to better model the ordinal nature of suicide risk labels. MentalBERT (Ji et al., 2021) demonstrates an additional improvement over SISMO, with a 2% increase in F. score.

RISE significantly outperforms all baselines including SASI (Sawhney et al., 2022b) for all coverages with a 3% better F. score on average while being able to assess risk upto 2.9x and 3.5x earlier for the CSSRS and CLPsych 2022 datasets respectively. The CSSRS and CLPsych datasets have an average timeline spanning 44 and 60 days for each user, users posting once in two days on average. Therefore, RISE can help identify suicide risk 30 to 40 days earlier in a real life scenario using just 15 to 20 days' posting history (7-10 posts) for a user on average, making timely intervention. This demonstrates the ability of RISE as a practical approach for early suicide risk assessment due to it's ability to abstain and use fewer time steps (posts) to produce state-of-the-art results.

6.2. Ablation Study

We contextualize the impact of various components of RISE in Table 3 with the help of an ablation experiment on the CSSRS Dataset (Gaur et al., 2019). All models with the exception of LSTM are run on a coverage $\mathcal{C}$ of 85%. Generally, augmenting LSTM (Hochreiter and Schmidhuber, 1997) with an abstention mechanism (Liu et al., 2019; Geifman and El-Yaniv, 2017) leads to an average of 5.5% improvement in F. score. Gambler's Loss (GL) (Liu et al., 2019) works better than Softmax Response (SR) (Geifman and El-Yaniv, 2017) as the abstention mechanism outperforming it by 5% on F. score while being 9% more robust when augmented to the vanilla LSTM mechanism. Next, we see an average improvement of 7% in F. score on replacing the vanilla LSTM with LSTM-IC, while being 2 times faster. A further improvement of 1% is observed with the introduction of hyperbolic geometry in the LSTM architecture as the hyperbolic space better models the innate power-law dynamics and hierarchies in online text streams (Sala et al., 2018).

As a result, our best performing model is a product of a better abstention mechanism, internal classifier's early exiting abilities combined with the superior ability of hyperbolic spaces to better model

Model	Gr. Precision	Gr. Recall	F. Score	Robustness	Fail-safe Rejects	EDER
Contextual CNN	0.65	0.52	0.59	-	-	-
SDM	0.61	0.54	0.57	-	-	-
Context BERT	0.63	0.57	0.60	-	-	-
LSTM	0.64	0.59	0.60	-	-	-
n-BiLSTM	0.65	0.60	0.62	-	-	-
SISMO	0.66	0.61	0.63	-	-	-
MentalBERT	0.65	0.62	0.65	-	-	-
SASI ($\mathcal{C}$ 100%)	0.67	0.62	0.66	0.48	-	-
SASI ($\mathcal{C}$ 85%)	0.69	0.65	0.67	0.61	0.83	-
SASI ($\mathcal{C}$ 50%)	0.71	0.69	0.70	**0.73**	0.65	-
RISE ($\mathcal{C}$ 100%)	0.70*	0.72*	0.71*	0.61*	-	2.8x
RISE ($\mathcal{C}$ 85%)	0.70*	0.72*	0.71*	0.67*	**0.84***	2.7x
RISE ($\mathcal{C}$ 50%)	**0.72***	**0.73***	**0.72***	**0.73***	0.77*	**2.9x**

(a) CSSRS Dataset

Model	Gr. Precision	Gr. Recall	F. Score	Robustness	Fail-safe Rejects	EDER
Contextual CNN	0.42	0.42	0.42	-	-	-
SDM	0.40	0.41	0.41	-	-	-
Context BERT	0.42	0.44	0.43	-	-	-
LSTM	0.47	0.44	0.43	-	-	-
n-BiLSTM	0.48	0.47	0.44	-	-	-
SISMO	0.49	0.47	0.45	-	-	-
MentalBERT	0.50	0.50	0.47	-	-	-
SASI ($\mathcal{C}$ 100%)	0.52	0.50	0.52	0.41	-	-
SASI ($\mathcal{C}$ 85%)	0.54	0.53	0.54	0.58	0.77	-
SASI ($\mathcal{C}$ 50%)	0.55	0.57	0.56	0.65	0.61	-
RISE ($\mathcal{C}$ 100%)	0.55*	0.54*	0.54*	0.55*	-	3.3x
RISE ($\mathcal{C}$ 85%)	0.57*	0.55*	0.56*	0.60*	**0.80***	3.3x
RISE ($\mathcal{C}$ 50%)	**0.58***	**0.59***	**0.59***	**0.69***	0.74*	**3.5x**

(b) CLPsych 2022 Dataset

Table 2: Performance comparison of RISE with other baseline classifiers. Bold shows the best result. * shows significant (p<0.01) improvement over SASI.

Model	Gr. Precision	Gr. Recall	F. Score	Robustness	Fail-safe Rejects	EDER
LSTM	0.64	0.59	0.60	-	-	-
LSTM w SR	0.65	0.62	0.63	0.55	0.58	-
LSTM w GL	0.68	0.68	0.68	0.64	0.69	-
LSTM-IC w SR	0.66	0.71	0.69	0.59	0.65	1.9x
HLSTM-IC w SR	0.69	0.69	0.69	0.60	0.64	1.8x
LSTM-IC w GL	0.69	0.71	0.70	**0.74**	0.77	2.5x
HLSTM-IC w GL (RISE)	**0.70***	**0.72***	**0.71***	0.67	**0.82***	**2.7x***

Table 3: Ablation study of RISE with different model components and geometries on the CSSRS Dataset (Gaur et al., 2019). Bold shows the best result. * shows significant (p<0.01) improvement over LSTM. GL stands for Gambler's Loss while SR stands for Softmax Response, both working as abstention mechanisms.

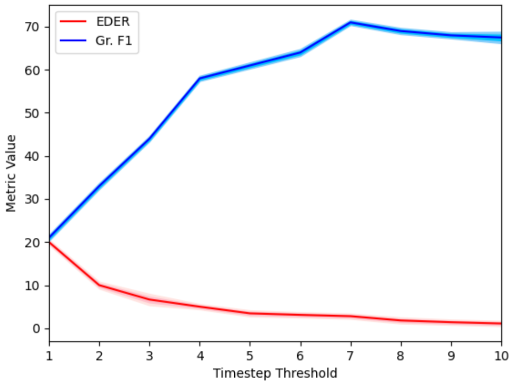


Figure 3: Impact of varying Time-step Threshold on model performance and efficiency.

online text stream (Sala et al., 2018).

6.3. Impact of Varying Time-step Threshold

We study the variation in RISE's performance and efficiency on introducing a Time-step threshold and varying it's value in Figure 3 using the CSSRS dataset. We restrict RISE to propagate through a minimum number of time steps before considering an early exit using threshold P. On gradually increasing P, we observe a significant improvement in performance upto to a certain optimal point suggesting how increasing context helps RISE in correctly classifying samples. This is followed by slight dip in performance accompanied by stagnation at the optimal value of 7. Although EDER rapidly decreases uptil this optimal point, RISE is still able to perform at par with state-of-the-art model like SISMO (Sawhney et al., 2021c) and SASI (Sawhney et al., 2022b) while being more efficient.

6.4. Qualitative Analysis

The effectiveness of RISE lies not just in its risk-averse nature due to the abstention mechanism but more importantly in its ability to predict correct labels with fewer time steps for a high risk user. We qualitative study the case of five users with their post samples from the CSSRS dataset as given in Figure 4. User I shows strong signs of high suicide risk severity early on. As a result, RISE is confidently able to correctly classify them using just seven posts (time steps). User II's low risk severity is clearly evident in their posts and is correctly classified by RISE in just nine time steps. The model makes an erroneous prediction for User III. However, RISE does not commit to this prediction, and abstains, indicating the requirement of a professional's immediate intervention. Even though RISE correctly classifies the suicide risk severity for User IV, it chooses to abstain. This is possibly due to RISE following a cautious approach due to the occurrence of phrases like "I do have a gun"

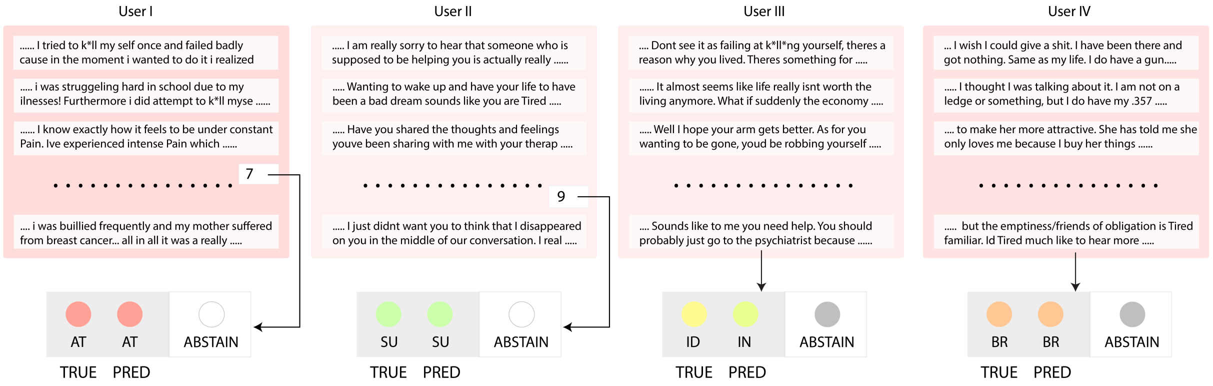

User I
...... I tried to k*ll my self once and failed badly cause in the moment i wanted to do it i realized
..... i was struggeling hard in school due to my ilnesses! Furthermore i did attempt to k*ll myse ......
...... I know exactly how it feels to be under constant Pain. Ive experienced intense Pain which ......
7
.... i was buillied frequently and my mother suffered from breast cancer... all in all it was a really .....
AT
AT
ABSTAIN
TRUE
PRED
User II
..... I am really sorry to hear that someone who is supposed to be helping you is actually really ......
..... Wanting to wake up and have your life to have been a bad dream sounds like you are Tired .....
..... Have you shared the thoughts and feelings youve been sharing with me with your therap .....
9
..... I just didnt want you to think that I disappeared on you in the middle of our conversation. I real .....
SU
SU
ABSTAIN
TRUE
PRED
User III
.... Dont see it as failing at k*ll*ng yourself, theres a reason why you lived. Theres something for .....
...... It almost seems like life really isnt worth the living anymore. What if suddenly the economy .....
..... Well I hope your arm gets better. As for you wanting to be gone, youd be robbing yourself .....
.... Sounds like to me you need help. You should probably just go to the psychiatrist because ......
ID
IN
ABSTAIN
TRUE
PRED
User IV
... I wish I could give a shit. I have been there and got nothing. Same as my life. I do have a gun.....
..... I thought I was talking about it. I am not on a ledge or something, but I do have my .357 .....
.... to make her more attractive. She has told me she only loves me because I buy her things ......
..... but the emptiness/friends of obligation is Tired familiar. Id Tired much like to hear more .....
BR
BR
ABSTAIN
TRUE
PRED

Figure 4: We show RISE can be used for efficient prioritization of users during suicide risk assessment with the help of the CSSRS dataset. For each user, we show the real labels next to predicted labels, while also indicating whether RISE refrained from making that prediction. We further demonstrate how RISE predicts correct samples early on without propagating through all time-steps.

repeatedly. There may be cases in which RISE confidently classifies a high risk user as a low risk user. Handling such cases is critical and is a current limitation of RISE.

7. Conclusion

In response to the pressing need for a robust solution for fine-grained suicide risk assessment on social media platforms, we introduce RISE, an innovative framework that integrates selective prioritization and early exit inference mechanism into existing deep learning-based risk assessment techniques. RISE embodies self-awareness by abstaining from making predictions when faced with uncertainty. It managed to out-perform current state-of-the-art suicide risk assessment models while being upto 3.5x faster. Through extensive quantitative evaluations conducted on real-world data, RISE demonstrated its effectiveness by successfully avoiding high-risk situations, abstaining from making upto 84% of incorrect predictions. Furthermore, we provide a detailed qualitative analysis highlighting the potential application of RISE within a human-in-the-loop framework, enabling timely and efficient responses from mental health experts.

8. Ethical Considerations

The research we present raises significant ethical considerations, particularly regarding the balance between privacy and effectiveness. Following the insights provided by (Coppersmith et al., 2018), we prioritize adherence to acceptable privacy practices to avoid coercion and intrusive treatment as outlined by (Fiesler and Proferes, 2018; Chancellor et al., 2019). The datasets used in this study are sourced from Reddit, a platform designed for anonymous posting. However, to ensure additional privacy safeguards, we employ automated de-identification techniques using named entity recognition (Zirikly et al., 2019) on the datasets. Furthermore, all examples utilized in this paper are anonymized, obfuscated, and paraphrased following the moderate disguise scheme proposed by (Bruckman, 2002) and (Benton et al., 2017).

Additionally, it is crucial to prevent overburdening clinicians and human moderators (Chancellor et al., 2019), considering challenges like "alarm fatigue" in healthcare, where excessive false positives can desensitize healthcare providers (Drew et al., 2014). We also acknowledge the subjective nature of suicidality (Keilp et al., 2012), where interpretations may vary among individuals on social media. We do not make any diagnostic claims but rather aim to prioritize users who should be evaluated first by medical professionals as part of a distributed human-in-the-loop framework (Andrade et al., 2018).

9. Limitations

We acknowledge the limitation of our work. First, our model was evaluated only on predominantly English datasets. The effectiveness of RISE may vary across different languages and cultural contexts.

We recognize that the analysed data may be influenced by demographic, expert annotator, and medium-specific biases (Hovy and Spruit, 2016). While our work aims to assist in the early detection of at-risk users and early intervention, it is essential to carefully plan and execute interventions to avoid counter-helpful outcomes, such as users migrating to fringe platforms, which can make providing assistance more challenging (Kumar et al., 2015).

To maintain user privacy, the annotation of user data is stored separately from the raw user data on protected servers, linked only through anonymous

IDs. Our objective is to develop an assistive tool for screening suicidal users based solely on observational capacity. However, we acknowledge that it is challenging to entirely prevent the misuse of technology, even when developed with good intentions (Hovy and Spruit, 2016).

10. Bibliographical References

Shivam Agarwal, Ramit Sawhney, Sanchit Ahuja, Ritesh Soun, and Sudheer Chava. 2022. Hyphen: Hyperbolic hawkes attention for text streams. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 620–627.

Norberto Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. 2018. Ethics and artificial intelligence: Suicide prevention on facebook. *Philosophy Technology*, 31:1–16.

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*, KDD '17, page 65–74, New York, NY, USA. Association for Computing Machinery.

Gary Bécigneul and Octavian-Eugen Ganea. 2018. Riemannian adaptive optimization methods. *CoRR*, abs/1810.00760.

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.

Anna D Broido and Aaron Clauset. 2019. Scale-free networks are rare. *Nature communications*, 10(1):1017.

Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the internet. *Ethics and Information Technology*, 4:217–231.

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.

CDC. 2021. Suicide. `https://www.cdc.gov/suicide/facts/index.html`.

Dave Chaffey. 2023. Global social media statistics research summary 2023. `https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/`.

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.

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*, 10:1178222618792860.

Kate Daine, Keith Hawton, Vinod Singaravelu, Anne Stewart, Sue Simkin, and Paul Montgomery. 2013. The power of the web: a systematic review of studies of the influence of the internet on self-harm and suicide in young people. *PloS one*, 8(10):e77555.

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.

Bart Desmet and Véronique Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Barbara Drew, Patricia Harris, Jessica Zègre-Hemsey, Tina Mammone, Daniel Schindler, Rebeca Salas-Boni, Yong Bai, Adelita Tinoco, Quan Ding, and Xiao Hu. 2014. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PLoS One*, 9:e110274.

Robert A Fahey, Jeremy Boo, and Michiko Ueda. 2020. Covariance in diurnal patterns of suicide-related expressions on twitter and recorded suicide deaths. *Social Science & Medicine*, 253:112960.

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

Xavier Gabaix. 2016. Power laws in economics: An introduction. *Journal of Economic Perspectives*, 30(1):185–206.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic neural networks. In *NeurIPS*, pages 5350–5360.

Muskan Garg. 2023. Mental health analysis in social media posts: A survey. *Archives of Computational Methods in Engineering*, pages 1–24.

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*, WWW '19, page 514–525, New York, NY, USA. Association for Computing Machinery.

Manas Gaur, Vamsi Aribandi, Amanuel Alambo, Ugur Kursuncu, Krishnaprasad Thirunarayan, Jonathan Beich, Jyotishman Pathak, and Amit Sheth. 2021. Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-ssrs. *PLOS ONE*, 16:e0250448.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4885–4894, Red Hook, NY, USA. Curran Associates Inc.

Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. 2019. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*.

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, Berlin, Germany. Association for Computational Linguistics.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 261–269, New York, NY, USA. Association for Computing Machinery.

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*.

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*, 8(1):214–226.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare.

John Keilp, Michael Grunebaum, Marianne Gorlyn, Simone LeBlanc, Ainsley Burke, Hanga Galfalvy, Maria Oquendo, and J. Mann. 2012. Suicidal ideation and the subjective aspects of depression. *Journal of affective disorders*, 140:75–81.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. 2020. Hyperbolic image embeddings. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages

1746–1751, Doha, Qatar. Association for Computational Linguistics.

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM Conference on Hypertext amp; Social Media*, HT '15, page 85–94, New York, NY, USA. Association for Computing Machinery.

Vasileios Lampos, Tijl De Bie, and Nello Cristianini. 2010. Flu detector - tracking epidemics on twitter. In *Machine Learning and Knowledge Discovery in Databases*, pages 599–602, Berlin, Heidelberg. Springer Berlin Heidelberg.

Victor Leiva and Ana Freire. 2017. Towards suicide prevention: early detection of depression on social media. In *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4*, pages 428–436. Springer.

Kaiyuan Liao, Yi Zhang, Xuancheng Ren, Qi Su, Xu Sun, and Bin He. 2021. A global past-future early exit method for accelerating inference of pre-trained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2013–2023, Online. Association for Computational Linguistics.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. 2019. Deep gamblers: Learning to abstain with portfolio theory. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Lauren McGillivray, Demee Rheinberger, Jessica Wang, Alexander Burnett, and Michelle Torok. 2022. Non-disclosing youth: a cross sectional study to understand why young people do not disclose suicidal thoughts to their mental health professional. *BMC Psychiatry*, 22.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Michael Paul and Mark Dredze. 2021. You are what you tweet: Analyzing twitter for public health. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):265–272.

Kelly Posner, Gregory Brown, Barbara Stanley, Kseniya Yershova, Maria Oquendo, Glenn Currier, Glenn Melvin, Laurence Greenhill, Sa Shen, and J. Mann. 2011. The columbia-suicide severity rating scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *The American journal of psychiatry*, 168:1266–77.

Kaushik Roy, Usha Lokala, Manas Gaur, and Amit Sheth. 2022. Tutorial: Neuro-symbolic ai for mental healthcare.

Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. 2018. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4457–4466. PMLR.

Ramit Sawhney, Shivam Agarwal, Atula Tejaswi Neerkaje, Nikolaos Aletras, Preslav Nakov, and Lucie Flek. 2022a. Towards suicide ideation detection through online conversational context. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1716–1727.

Ramit Sawhney, Shivam Agarwal, Megh Thakkar, Arnav Wadhwa, and Rajiv Ratn Shah. 2021a. Hyperbolic online time stream modeling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1682–1686.

Ramit Sawhney, Shivam Agarwal, Megh Thakkar, Arnav Wadhwa, and Rajiv Ratn Shah. 2021b. *Hyperbolic Online Time Stream Modeling*, page 1682–1686. Association for Computing Machinery, New York, NY, USA.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021c. Towards ordinal suicide ideation detection on social media. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 22–30, New York, NY, USA. Association for Computing Machinery.

Ramit Sawhney, Atula Tejaswi Neerkaje, and Manas Gaur. 2022b. A risk-averse mechanism for suicidality assessment on social media. *Association for Computational Linguistics 2022 (ACL 2022)*.

Dr S Smys and Dr Jennifer S Raj. 2021. Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. *Journal of Trends in Computer Science and Smart Technology*, 3(1):24–39.

Kim Stene-Larsen and Anne Reneflot. 2019. Contact with primary and mental health care prior to suicide: A systematic review of the literature from 2000 to 2017. *Scandinavian Journal of Public Health*, 47(1):9–17. PMID: 29207932.

Tianxiang Sun, Xiangyang Liu, Wei Zhu, Zhichao Geng, Lingling Wu, Yilong He, Yuan Ni, Guotong Xie, Xuanjing Huang, and Xipeng Qiu. 2022. A simple hash-based early exiting approach for language understanding and generation.

Tianxiang Sun, Yunhua Zhou, Xiangyang Liu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. 2021. Early exiting with ensemble internal classifiers.

Alexandru Tifrea, Gary Becigneul, and Octavian-Eugen Ganea. 2019. Poincare glove: Hyperbolic word embeddings. In *International Conference on Learning Representations*.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

Teun Adrianus Van Dijk. 1977. *Text and context: Explorations in the semantics and pragmatics of discourse*. Longman London.

WHO. 2021. Suicide. `https://www.who.int/news-room/fact-sheets/detail/suicide`.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. 2020. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2246–2251, Online. Association for Computational Linguistics.

Yunxiang Zhang and Zhuyi Rao. 2020. n-bilstm: Bilstm with n-gram features for text classification. *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*, pages 1056–1059.

Xiaojun Zhao, Pengjian Shang, and Yulei Pang. 2010. Power law and stretched exponential effects of extreme events in chinese stock markets. *Fluctuation and Noise Letters*.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. 2020. Bert loses patience: Fast and robust inference with early exit.

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.

Liu Ziyin, Blair Chen, Ru Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. 2020. Learning not to learn in the presence of noisy labels. *arXiv preprint arXiv:2002.06541*.