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

Motivation

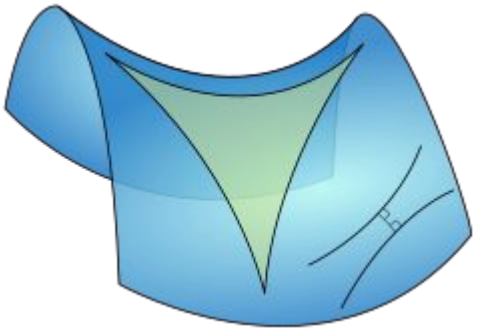
- Suicide is a serious public health issue, but it is preventable with timely intervention.
- Due to increase in the number of individuals sharing suicidal thoughts online, utilising advance Natural Language Processing techniques to build automated systems for risk assessment is a viable alternative.
- Existing systems are prone to incorrectly predicting risk severity and have no early detection mechanisms.

Introduction

- We introduce **RISE** to effectively gauge suicide risk of a user engaging in online forums through their post history.
- We model the hyperbolic nature of online text streams using a better suited geometry that captures the powerlaw dynamics in social media texts
- We formulate **RISE**, a novel risk-averse mechanism for early detection of suicide risk by ensembling Hyperbolic Internal Classifiers equipped with an abstention mechanism and early-exit inference capabilities
- Through ablative, qualitative and quantitative experiments, we demonstrate the ability of RISE as a robust and efficient approach for early detection of suicide risk using online text streams.

RISE

- **RISE** operates in the hyperbolic geometry instead of the standard euclidean geometry to effectively capture tree-like hierarchical structures that online text-streams exhibit.
- **RISE** uses **BERT** to embed social media posts of users, following which these embeddings are projected to the hyperbolic space using exponential map.



RISE (Continued)

- **RISE** uses Hyperbolic-LSTM with internal classifiers (MLP at every time-step) to enable early exit decision in case a user exhibits suicide risk early on.
- This makes it possible for **RISE** to make a confident predictions with fewer number of posts (user history).

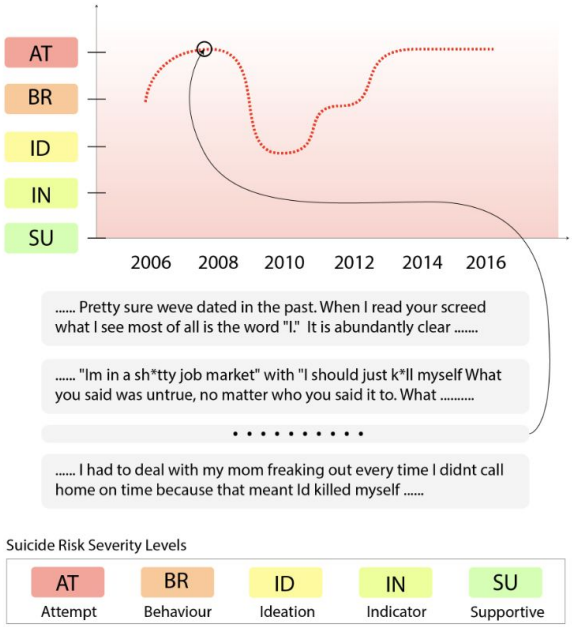
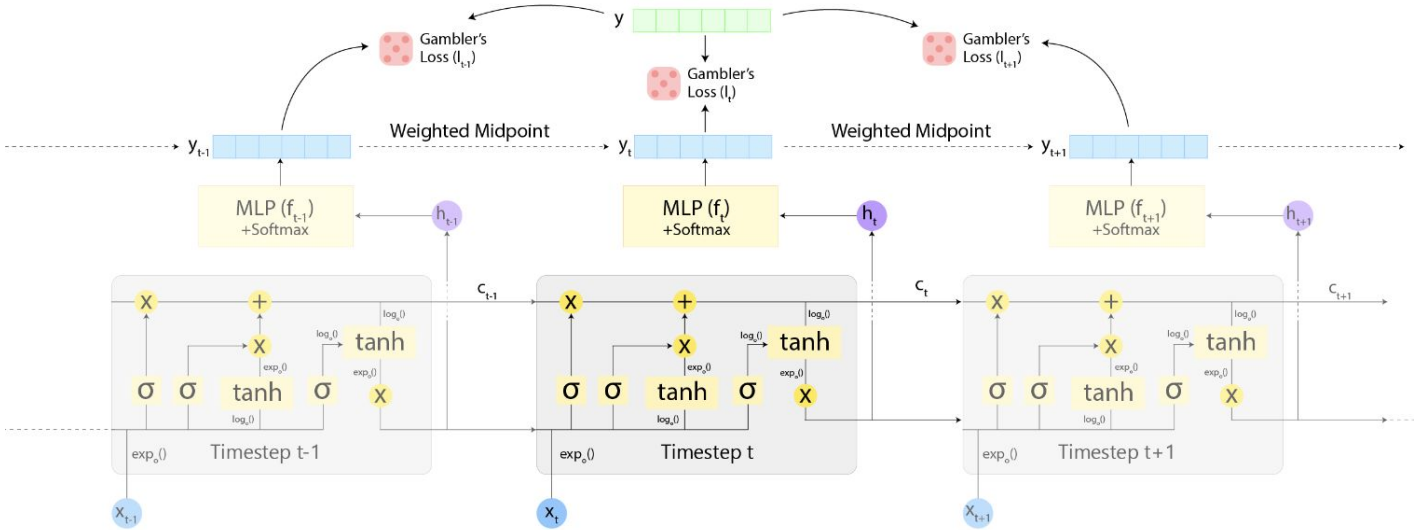


Figure 1: We visualize the suicide risk severity of a sample of Reddit posts for a user from the CSSRS Dataset with “Attempt (AT)” risk severity and plot it over time.

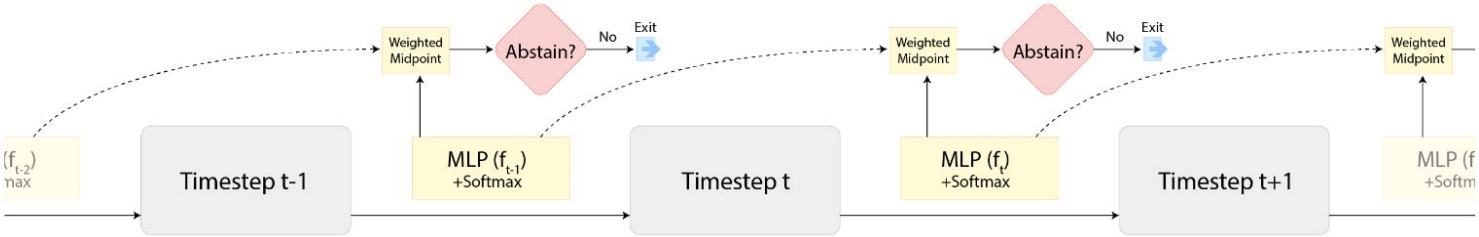
RISE (Continued)

- 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 by augmenting the label space with an option to abstain.

Training


y
Gambler's Loss (l_{t-1})
Gambler's Loss (l_t)
Gambler's Loss (l_{t+1})
Weighted Midpoint
Weighted Midpoint
y_{t-1}
y_t
y_{t+1}
MLP (f_{t-1}) +Softmax
MLP (f_t) +Softmax
MLP (f_{t+1}) +Softmax
h_{t-1}
h_t
h_{t+1}
c_{t-1}
c_t
c_{t+1}
σ
tanh
Timestep t-1
Timestep t
Timestep t+1
x_{t-1}
x_t
x_{t+1}

Inference


Weighted Midpoint
Abstain?
No
Exit
Weighted Midpoint
Abstain?
No
Exit
Weighted Midpoint
Timestep t-1
MLP (f_t-1) +Softmax
Timestep t
MLP (f_t) +Softmax
Timestep t+1

Evaluation Metrics

- We use **graded Precision, Recall and F1-score** to evaluate the classification performance of RISE.
- We also introduce some new metrics to evaluate the robustness, fail-safe nature and early exiting nature of RISE:
    - **Fail-safe rejects**: It is given as the ratio of number of incorrect prediction by the number of samples abstained
    - **Robustness**: Robustness is quantified as the fraction of samples correctly classified or abstained
    - **Early Detection Efficiency Ratio**: 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.

Performance Comparison

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.

Ablation Study

- 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 online text streams.

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.

Impact of Varying Time-step Threshold

- 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 models.

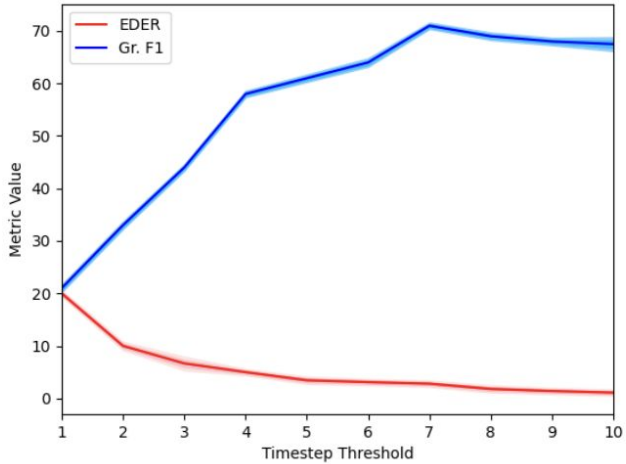


Figure 3: Impact of varying Time-step Threshold on model performance and efficiency.

Qualitative Analysis

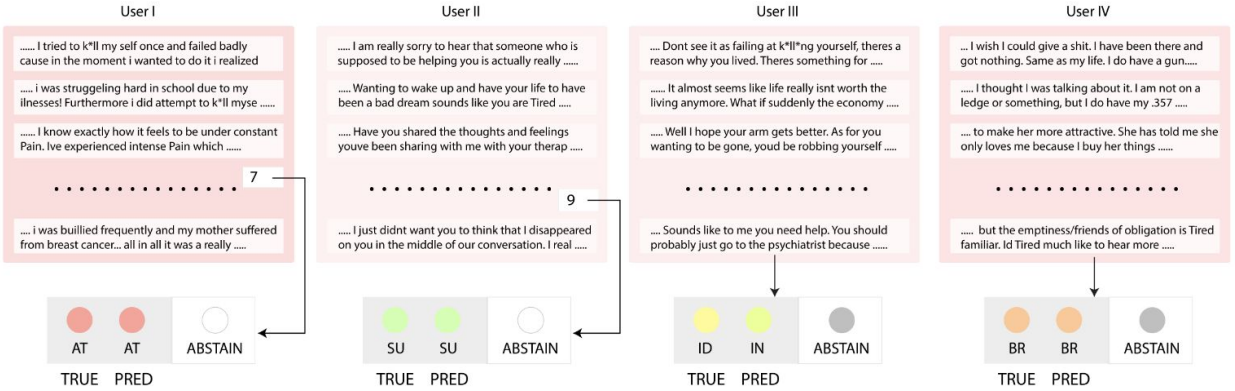


Figure 4: We show RISE can be used for efficient prioritization of users during suicide risk assessment with the help of the CSSRS dataset. For each user, we show the real labels next to predicted labels, while also indicating whether RISE refrained from making that prediction. We further demonstrate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

Conclusion

- we introduce RISE, an innovative framework that integrates selective prioritization and early exit inference mechanism into existing deep learning-based risk assessment techniques.
- It managed to out-perform current state-of-the-art suicide risk assessment models while being upto 3.5x faster.
- Through extensive quantitative, qualitative and ablative evaluations conducted on real-world data, RISE demonstrated its effectiveness as a viable solution.

Thank you!