AI-Powered Suicide Risk Detection in Social Media Posts Using Deep Learning Models

Saraf Pratyaksh,
School of Social Science,
NTU

Dr. Ong Chi Wei,
School of Chemistry,
Chemical Engineering and
Biotechnology, NTU

Prof Cyrus Ho,
Yong Loo Lin School of
Medicine, NUS

Abstract - Even though mental health awareness has grown reasonably over the decade, stigma and privacy concerns still deter many from seeking help, increasing the risk of untreated conditions escalating to suicidal tendencies. The rise of social media platforms like Reddit, which support anonymity, has enabled various individuals to share their emotions freely, providing a valuable source for data analysis. This study leverages this resource to develop a deep learning model to improve the early detection and prevention of suicide risks among users by quantifying the risks based on linguistic patterns, sentiment shifts and behavioural cues.

We trained our model using the Suicidal Ideation Reddit Dataset (Shukla, 2024) [1], containing textual posts scraped from Twitter and Reddit, binary labelled as 'Suicide' or 'Non-Suicide.' Preprocessing involved text cleaning, tokenization, and feature extraction. Although Long Short-Term Memory (LSTM) frameworks served as a baseline, BERT and XLNet, with their transformer frameworks, provided progressively higher fidelity in capturing contextual meaning. The fine-tuned BERT model achieved a 90% accuracy, with strong ROC-AUC metrics and F1-score.

An interactive interface was further developed using Streamlit to demonstrate real-world applicability, enabling users to input text and receive a probabilistic risk index together with a corresponding risk category. This tool showcases how AI-driven mental health interventions can aid researchers and professionals in proactive response efforts. Future works could improve model interpretability and integrate multi-modal data to enhance detection accuracy and real-world applicability.

1. INTRODUCTION

Suicide remains a leading cause of death worldwide, with nearly **7,20,000** deaths reported annually [2]. Young adults often struggle with isolation and mental distress, yet social stigma and privacy concerns deter many from reaching out and seeking help. Social media platforms, especially those that support anonymity, such as Reddit, often become comfortable spaces for such individuals to openly express their thoughts and feelings, which might otherwise remain hidden. Thus, discussions within forums such as r/SuicideWatch on Reddit can provide a valuable resource when analysing thought patterns in individuals to assess their elevated risk.

This study investigates the potential of leveraging deep learning models to reliably ascertain suicidal intent within user-generated text. By training and evaluating models on a large, labelled dataset of Reddit and Twitter posts, we aim to develop a model capable of flagging high-risk content. The broader goal is to integrate this system into outreach tools that assist psychologists, professionals, and support organizations in early detection efforts.

2. LITERATURE REVIEW

Early work on suicide risk detection focused on lexicon-based filtering and classical machine learning. Coppersmith et al. (2015) applied predefined keyword lists and sentiment lexicons to Twitter data, achieving moderate recall but high false positives in the face of slang and irony [3]. Moving beyond static features, Gaur et al. (2019) employed an LSTM network on Reddit posts. By feeding text embeddings into a 128-unit LSTM later, they reported a 78% accuracy, outperforming SVM and logistic regression baselines, and demonstrated that longer sequences (up to 200 tokens) better captured evolving emotional cues [4].

The advent of transformer models marked a significant advance. Devlin et al. (2019) introduced BERT, which learns bidirectional context through masked-language modelling [5]. Building on this, Chancellor et al. (2021) fine-tuned BERT on combined Reddit and Twitter suicide data, achieving 88% accuracy and 0.92 ROC-AUC. Their analysis also showed that attention weights corresponded to human-annotated distress markers [6].

Meanwhile, Yang et al. (2019) proposed XLNet, a permutation-based transformer that avoids making biases. When fine-tuned for mental health classification, XLNet outperformed BERT by 2-3% accuracy and ROC-AUC, highlighting a trade-off

between marginal gains and increased computational cost [7].

Despite these improvements, most studies treat suicide detection as a binary classification problem, with limited integration of interpretability or continuous risk scoring. Ribeiro et al. (2016) developed LIME to produce local explanations for black-box models [8], yet few suicide-risk systems have adopted it. Moreover, existing research typically lacks real-time deployment and multilingual support, limiting practical applicability in diverse settings.

Our work addresses these gaps by (1) mapping outputs to a **five-level Suicide Risk Index**, (2) using mean squared error loss to encourage the model to learn nuanced risk probabilities, and (3) deploying **fine-tuned BERT encoder** within a **Streamlit** interface [9] equipped with automated translation and language detection. This combination enhances interpretability, supports global user input, and bridges the divide between prototype performance and real-world utility.

3. METHODOLOGY

We utilized the '**Suicide Ideation Reddit Dataset**' from the IEEE Data Port, a publicly available dataset with textual posts scraped from Reddit and Twitter, binarily labelled as '**Suicide**' or '**Non-Suicide**' [1].

To benchmark performance, we developed three different neural network architectures. Each model was designed to analyze textual input data and output suicide risk probabilities between 0 and 1.

The first was a simple LSTM, serving as our baseline model. It included an embedding layer to convert words into vector representations, followed by a single 256-unit hidden LSTM layer. A final sigmoid function was later applied for binary classification. The model was trained using the binary cross-entropy loss function and the Adam optimizer, enabling it to learn directly from the word embeddings and temporal patterns in the text data.

The second model employed an XLNet-base, a transformer model pre-trained on a large text corpora. The model was extended by adding a dropout layer and a fully connected sigmoid output layer. Each layer of the XLNet was fine-tuned during training to allow the model to better capture the nuances of suicidal vs non-suicidal language in our dataset. The training was performed over 12 epochs using binary cross-entropy loss and the Adam optimizer.

The final model was based BERT-base (uncased), similarly fine-tuned on the dataset. A custom

classification head was built on top of the transformer encoder, consisting of two

fully connected layers with ReLU activations and a dropout, followed by a sigmoid output layer. The model was trained using mean squared error (MSE) loss to emphasize the regression-like interpretation of outputs like suicide risk probabilities rather than discrete classes.

Figure 1. Training the Fine-Tuned BERT Model

Figure 2. BERT Fine-Tuning Code Snippet

All three models were trained on the labelled Reddit and Twitter posts. Texts were pre-processed to remove special characters and a consistent train-test split of 80-20 was applied during training. Tokenization was handled using the respective pre-trained encoders for BERT and XLNet, with a maximum sequence length of 128.

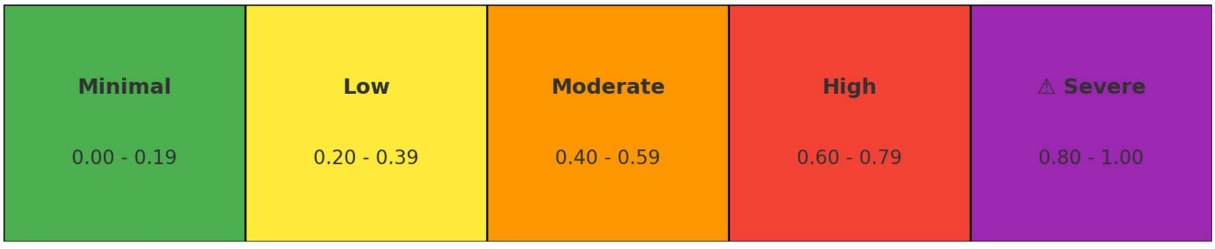

Figure 3. Suicide Risk Index Classification

Table 1. Distribution of Training and Test Posts

Split	*#Suicide Posts*	*#Non-Suicide Posts*	*Total*
Train	18,322	19,385	37,707
Test	4581	4847	9428

Model performance was evaluated on the test set using multiple metrics: accuracy, precision, recall, F1-score, and ROC-AUC. The predicted probabilities were interpreted as continuous risk scores, beyond binary evaluation, which were later categorized into five risk levels using our custom Suicide Risk Index. This index translated the risk scores into the corresponding risk bands-Minimal, Low, Moderate, High, and Severe.

Table 2. Suicide Risk Category Descriptions

CATEGORY	*DESCRIPTION*
Minimal Risk (0.00-0.19)	No significant indications of suicide risk.
Low Risk (0.20 - 0.39)	Some distress, but no strong indications of suicidality.
Moderate Risk (0.40 - 0.59)	Shows distress or mild suicidal ideation.
High Risk (0.60 - 0.79)	Strong signs of suicidal ideation, requires attention.
Severe Risk (0.80 - 1.00)	Critical risk may require immediate intervention.

Comparing these models under identical training and evaluation conditions helped us understand the trade-offs between classical Recurrent Neural Networks (RNNs) and modern transformer architectures, as well as the added advantage of fine-tuning Large Language models (LLMs) for mental-health related Natural Language Processing (NLP) tasks.

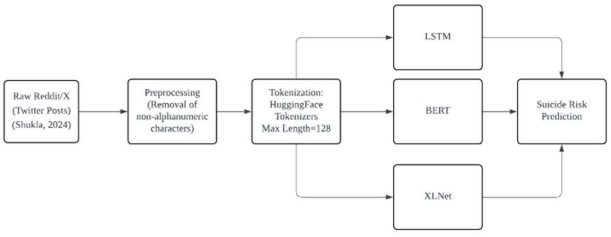

Figure 4. Model Architecture Pipeline

4. TECHNOLOGIES & IMPLEMENTATION

The models were developed entirely in Python, utilizing a combination of modern deep learning techniques and data processing libraries. PyTorch [10] was used to carry out core model development and training, given its flexibility to define custom neural architectures and manage GPU-accelerated training workflows. We relied on the HuggingFace Transformers library [11] for working with pretrained transformer models like BERT and XLNet, which streamlined access to tokenizers, model weights, and associated libraries.

Data preprocessing was conducted using pandas [12] for tabular operations and regular expressions for raw input texts, including the removal of non-alphanumeric characters. Performance metrics – accuracy, precision, recall, F1-score, and ROC-AUC – were computed using scikit-learn [13], which facilitated standardized model evaluation across experiments.

We utilized CUDA-enabled GPUs [14] to accelerate training, allowing parallelized matrix operations and efficient backpropagation for large-scale transformer models. This helped significantly reduce training time, especially during fine-tuning phases.

In an attempt to make our final BERT-based model more interactive and accessible, we deployed it through a web interface built with StreamLit. This lightweight frontend allows users to input text and receive a real-time risk assessment from our fine-tuned model. To support multilingual inputs, we integrated googletrans [15] for automated translation and langdetect [16] to identify input text. These components enabled our tool to generalize to non-English texts by dynamically translating user inputs into English before inference.

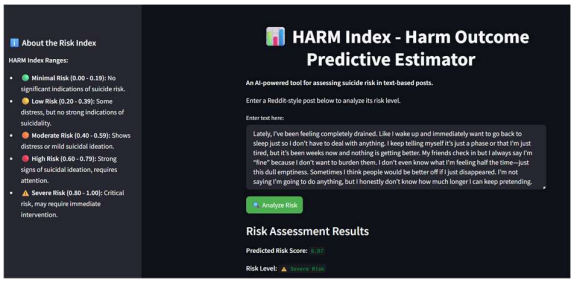

Figure 5. Streamlit UI Demonstrating Risk Prediction

(Screenshot of the deployed Streamlit UI showing a sample AI-generated input text and its corresponding predicted suicide risk score and category)

5. RESULTS

A clear performance improvement was observed across the three models, as the architectural complexity and use of pretraining increased. The LSTM model, lacking any form of transfer learning, achieved a **test accuracy of 72%** and an **ROC-AUC of 0.77**, indicating that it could detect basic sequential patterns in text but struggled to capture deeper contextual meanings.

The performance was substantially improved with XLNet, a transformer pretrained on extensive language data. After fine-tuning on our dataset, the model achieved an **85% accuracy** and **ROC-AC of 0.92**. This reflects the model's ability to use bidirectional context and long-range dependencies through its permutation-based self-attention mechanism.

The fine-tuned BERT-base was our highest-performing model, enhanced with a multilayer perceptron thread. It achieved a **90% accuracy** and an **ROC-AUC of 0.96**, substantially outperforming the LSTM and XLNet baselines. These results highlight BERT's effectiveness in adapting to domain-specific classifications, particularly those involving subtle emotional and psychological cues in language.

Table 3. Model Performance Comparison

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*ROC-AUC*
LSTM	0.72	0.84	0.52	0.64	0.77
XLNet	0.85	0.87	0.80	0.84	0.92
BERT	0.90	0.90	0.90	0.90	0.96

6. DISCUSSION

The strong performance of transformer-based models, particularly BERT, highlights the growing potential of integrating AI tools in supporting mental health. Achieving a high accuracy of 90% and an ROC-AUC score of 0.96, our fine-tuned BERT model not only demonstrates technical efficacy but also its potential real-world applications, where timely detection of suicide intent could prompt early intervention. These models offer scalable support systems, especially in settings with limited human moderation or where individuals may be reluctant to seek help directly.

The deployment of such tools opens promising avenues for integration into mental health platforms, digital well-being applications, or content moderation pipelines. For example, the risk probabilities could be utilized to triage flagged content, enabling more efficient allocation of human resources in crisis response teams. Additionally, the model's ability to quantify risk on a continuous scale, rather than a binary classification, supports a more nuanced understanding of user intent and emotional state.

However, translating these capabilities into responsible and effective systems also introduces important ethical, interpretability and robustness challenges. Given that mental health data is deeply personal, automated prediction tools must be designed in a way that safeguards privacy, informed consent, and responsible usage. Furthermore, misclassifications carry significant consequences – a false negative may delay intervention, while a false positive could result in unnecessary escalation or stigmatization [17].

The need for transparency is equally crucial in model decision-making. Black-box predictions can undermine trust, especially in clinical or crisis support settings where explainability is vital. Future improvements could integrate interpretability methods such as attention maps, LIME, or SHAP to provide actionable insights into why certain posts are flagged as high risk.

Furthermore, the current models operate solely on textual input. This approach might overlook potentially valuable signals from user metadata, behavioural patterns, or temporal posting trends. Incorporating such multimodal data could improve the model's sensitivity to evolving mental states and reduce overreliance on surface-level linguistic cues.

In summary, our results demonstrate the promise of transformer models in suicide risk detection, with immediate utility in scalable screening tools. However, realizing their full potential requires thoughtful design, grounded not just in technical performance, but in ethical responsibility, interpretability, and human-centered deployment.

ACKNOWLEDGMENT

I would like to extend my deepest gratitude to my Professor Ong Chi Wei for allowing me to join this project and giving me valuable guidance throughout the course of this project. I have had an inspiring time during this project.

Special gratitude to Professor Cyrus Ho for his constant support through the project.

I would like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project.

REFERENCES

[1] Shukla, S. (2024). *Suicidal ideation Reddit dataset*. IEEE DataPort. https://doi.org/10.21227/6fy2-tk76

[2] World Health Organization. (2025). *Suicide*. https://www.who.int/news-room/fact-sheets/detail/suicide

[3] Coppersmith, G., Dredze, M., & Harman, C. (2015). Quantifying mental health signals in Twitter. *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, 51–60.

[4] Gaur, M., Kumaraguru, P., & Sureka, A. (2019). Detecting suicidal ideation in social media: A lexical analysis of Reddit data. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 96–103). IEEE.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*. https://arxiv.org/abs/1810.04805

[6] Chancellor, S., Brown, J., Pater, J., Daumé III, H., & De Choudhury, M. (2021). Quantifying and predicting mental illness severity in online posts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)* (pp. 1–13). ACM.

[7] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint*. https://arxiv.org/abs/1906.08237

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

[9] Streamlit. (n.d.). *The fastest way to build and share data apps*. https://streamlit.io/

[10] PyTorch. (n.d.). *An open-source machine learning library for Python*. https://pytorch.org/

[11] HuggingFace. (n.d.). *Transformers: State-of-the-art NLP for PyTorch and TensorFlow 2.0*. https://huggingface.co/transformers/

[12] pandas Development Team. (n.d.). *pandas: Python data analysis library*. https://pandas.pydata.org/

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

[14] NVIDIA. (n.d.). *CUDA toolkit: GPU-accelerated computing*. https://developer.nvidia.com/cuda-toolkit

[15] Googletrans Developers. (n.d.). *googletrans (unofficial): Free Google Translate API for Python*. https://py-googletrans.readthedocs.io/

[16] Langdetect Developers. (n.d.). *langdetect: Language detection library ported from Google's language-detection*. https://pypi.org/project/langdetect/

[17] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679