DiGrI: Distorted Greedy Approach for Human-Assisted Online Suicide Ideation Detection

Usman Naseem
usman.naseem@mq.edu.au
Macquarie University
Sydney, New South Wales, Australia

Liang Hu*
lianghu@tongji.edu.cn
Tongji University
Shanghai, China

Qi Zhang
zhangqi_cs@tongji.edu.cn
Tongji University
Shanghai, China

Shoujin Wang
Shoujin.Wang@uts.edu.au
University of Technology Sydney
Sydney, New South Wales, Australia

Shoaib Jameel
M.S.Jameel@southampton.ac.uk
University of Southampton
Southampton, United Kingdom

Abstract

User-generated content on social media platforms provides a valuable resource for developing automated computational methods to detect mental health issues online leading to suicidal thoughts automatically. Although current fully automated methods show promise, they may produce uncertain predictions, leading to flawed conclusions. To address this, we propose a novel model called DiGrI, or **Di**storted **Gr**eedy Approach for Human-Assisted Online Suicide **I**deation Detection, which reformulates suicide ideation assessment as a selective, prioritized prediction problem. The model incorporates a novel multi-classifier distorted greedy model that is optimized to operate under various levels of automation and abstains from making uncertain predictions with theoretical guarantees. Our results show that DiGrI outperforms strong comparative models including large language models in detecting mental health issues on a publicly available Reddit dataset. We discuss the empirical and practical implications, including the ethical considerations of using DiGrI for online automatic suicide ideation detection involving humans, if it were to be translated for use in clinical and public health practice.

CCS Concepts

• **Applied computing** → **Document analysis**.

Keywords

Social Media Analysis; Mental Health Detection; Deep Learning; Suicide Ideation

ACM Reference Format:
Usman Naseem, Liang Hu, Qi Zhang, Shoujin Wang, and Shoaib Jameel. 2025. DiGrI: Distorted Greedy Approach for Human-Assisted Online Suicide Ideation Detection. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3696410.3714529

*Corresponding Author

WWW '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1274-6/25/04
https://doi.org/10.1145/3696410.3714529

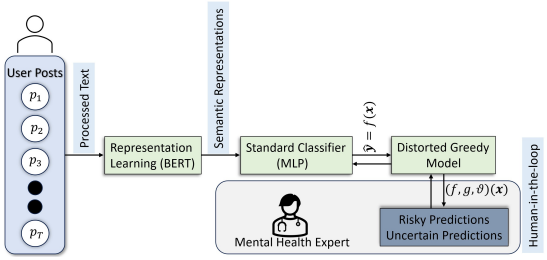

Figure 1: The end-to-end pipeline for online suicide ideation detection involves utilizing DiGrI to evaluate posts, providing predicted risk levels alongside corresponding certainty scores. In the context of a human-assisted approach, these predictions are categorized into distinct risk levels from High risk to completely uncertain. DiGrI strategically prioritizes uncertain and high-risk predictions, flagging them for review by mental health experts and ensuring that uncertain cases receive prompt attention and intervention. Our key contribution lies in developing the distorted greedy model.

1 Introduction

According to the World Health Organisation, suicide ideation is defined as, *"Thoughts, ideas, or ruminations about the possibility of ending one's life, ranging from thinking that one would be better off dead to the formulation of elaborate plans."* Each year, on average, almost 5000 people die of suicide in England and Wales [4]. This number has increased in recent years due to various factors, including the COVID-19 pandemic, social unrest, and economic inequality [29]. Several factors can cause these conditions, including genetics, brain chemistry, life experiences, and trauma. Unfortunately, three out of four people diagnosed with mental disorders do not receive treatment, which is alarming given the strong correlation between mental disorders and suicidal intentions [10, 14, 16, 46]. Studies reveal that approximately 900,000 individuals globally commit suicide every year[1] which is alarming. This paper focuses on automatic suicide prediction in a data-driven way, a topic that has garnered significant attention in recent literature [37].

Recent studies have revealed that individuals suffering from mental illnesses such as depression and harbouring suicidal thoughts often vent about personal issues on social media platforms instead

[1]https://www.who.int/news-room/fact-sheets/detail/suicide

of seeking professional help [21, 45, 50]. The reasons for this phenomenon range from inadequate funding to cover medical expenses to a desire to avoid social stigma and a lack of perceived justification for seeking medical help [48]. Early detection and intervention are crucial for improving mental health outcomes, but traditional methods of suicide ideation assessment are often time-consuming and expensive [39]. Consequently, there has been increasing interest in leveraging automated machine learning models and natural language processing (NLP) to identify warning signs for suicidal intention through user-generated social media posts to mitigate suicidal risk [9, 18, 37, 40, 50].

Even machine learning models cannot be fully relied upon, especially for tasks such as automated suicide prediction, as their performance may not always be guaranteed [5, 39]. This highlights the potential benefits of incorporating human feedback into the automated learning process to enhance the performance of computational models. Automated learning intertwined with human feedback is advantageous since humans possess the ability to understand and reason about complex problems, making sound judgements even under uncertain conditions. By incorporating human feedback into the machine learning process, it is possible to create more accurate and robust models than those trained solely on data. However, the key challenge lies in preventing humans from becoming overwhelmed by inaccurate predictions generated by the learning framework.

To address the above challenge, human involvement can be employed, where human experts are consulted to check the correctness of the model's predictions. This manual involvement can be used to refine the model's training data and improve its overall performance. Additionally, human experts can be involved in the decision-making process, ensuring that the model's predictions are interpreted and acted upon judiciously. By combining the strengths of automated machine learning and human expertise, it is possible to develop more effective and reliable systems for identifying individuals at risk of suicide. This can lead to timely interventions and improved mental health outcomes for those most in need [37].

Recently, Sawhney et al. [39] presented the SASI: Suicidality Assessment on Social MedIa model that improves automated suicide ideation prediction by factoring in uncertainty. The SASI model is a risk-averse and self-aware transformer-based model that refrains from making decisions when uncertain about the instance. The model can accurately predict suicide risk and pass uncertain cases to human experts. The authors found that their model is cautious and does not make incorrect predictions in 83% of cases, thereby improving reliability. In another recent study [5], the authors proposed a new framework for training machine learning classifiers to operate under different levels of automation. The authors argued that most supervised learning models are trained for full automation, but their predictions are sometimes worse than those by human experts in some specific instances. The authors developed a deterministic distorted greedy model for selecting the instances to be labelled by a human expert. The model works by iteratively adding the instance that maximizes the expected improvement in the classifier's performance. The model also takes into account the cost of labelling each instance by human experts.

Figure 1 illustrates the typical incorporation of the human involvement for suicide risk prediction. Mental health professionals with specialized expertise in suicide risk assessment act as the human experts involved. Chronologically arranged user posts are fed into a standard text classification model that estimates the severity of suicidal ideation in the individuals. The model then automatically categorizes these predicted levels into risk levels such as "High", "Moderate", or "Low". This categorization can be further refined automatically using either predefined threshold values [37] or, as our novel work proposes, by jointly fine-tuning the model parameters with the standard classifier, eliminating the need for static thresholds. Cases where the model is uncertain about its predictions (indicated by a newly created class) or where the model estimates high risk are automatically flagged as high-priority instances and routed to the health experts for further evaluation, along with the certainty score. This prioritization allows health experts to promptly review cases with high uncertainty, ensuring timely interventions for those in greatest need.

Contributions: We present a novel human-assisted computational model (`DiGrI`) that facilitates automated suicide risk prediction while minimizing the cognitive burden on human experts by prioritizing uncertain predictions to avert critical errors. Unlike prior works that rely on heuristic models with thresholds including sending all cases to human experts, we replace the heuristic threshold-based selection mechanism with a theoretically grounded selection mechanism that provides strong performance guarantees. We integrate a novel distorted greedy approximation technique with multiple classifiers as the new abstain function. The role of the abstain function is to inform the standard classifier whether to make an automatic prediction or not. We use trained classifiers to assign weights to a sample, enabling us to assess the proximity to human annotations. The greedy approximation technique provides a more nuanced evaluation of the model's uncertainty. The objective is then to maximize the evaluation metric of correct annotations. The greedy classifier selects samples for human review, iteratively selecting the sample with the highest expected improvement in the evaluation metric. It ensures that the model selects the most informative samples first, minimizing the cognitive load on human experts. Our experimental results demonstrate a substantial improvement over the previous best baseline model on a publicly available Reddit dataset.

2 Related Work

At the onset of automated mental health issue detection research, researchers employed a simplistic approach of feeding keywords to computers to identify posts potentially indicative of mental health issues [17]. For instance, researchers might use a list of keywords such as "depressed", "suicidal", and "self-harm" to flag posts that could be concerning. However, this approach was limited because it did not account for the context of the post or the user's personal information [50]. As an illustration, a post that says *"I'm so depressed"* might be a cry for help, but it could also be a sarcastic or ironic comment. Similarly, a post that says *"I'm going to kill myself"* might be a suicide threat, but it could also be a metaphor.

To address the limitations above, researchers developed more advanced models that could analyze various emotional states, multiple language styles, and peer-to-peer exchanges [6]. This allowed researchers to consider the effect and seriousness of the statements

used, creating more accurate detection models. One example is the Bidirectional Encoder Representations from Transformers (BERT) model [8], a pre-trained language model that can be fine-tuned for various tasks, including mental health detection [7]. BERT can learn the meaning of words and phrases in the context of a sentence, allowing it to better understand the sentiment and meaning of social media posts ([50]). Another popular example, before BERT, is the Long Short-Term Memory (LSTM) model [11], a type of neural network that can learn long-term dependencies. LSTMs are well-suited for tasks such as sentiment analysis and text classification, making them ideal for mental health detection. In these automated methods, there is still a risk of false positives and false negatives, and it is important to involve human experts during decision-making [37].

Given the serious nature of mental health issues and the unpredictability of human nature, the suicide assessment must be handled with care [39]. Therefore, it is essential to incorporate a human professional layer into these models to ensure that no lives are lost due to technological error. There are two additional reasons why this is important. First, Balazadeh Meresht et al. [2] found that the collaboration of human professionals and machines in Reinforcement Learning Agents enhances the performance of the algorithm. Second, in [5] the authors demonstrated that classification with human involvement under different levels of automation outperforms full automation levels.

There have been significant advancements in human-machine collaboration for suicidal thoughts detection [2, 5, 39, 42, 43]. For example, Sawhney et al. [39] introduced the Suicidality Assessment on Social Media (SASI) model, which measures the uncertainty of the machine learning model by expanding the cardinality of the label space. If the model is too uncertain, it refrains from predicting and asks a human professional to make the final decision. The preset data coverage parameter controls how often the human is allowed to intervene. The optimal level of automation has not yet been determined due to the preset variable. Balazadeh Meresht et al. [2] proposed using a two-layer Markov decision process and a reinforcement learning algorithm to find the optimal level of automation. In [5], the authors developed a distorted greedy model that learns the fundamental relationship between data and its corresponding machine and human error. This model incorporates the effect of machine and human error into the calculation. Overall, research on human-machine collaboration for suicide ideation detection is still in its early stages, but it has the potential to revolutionize the way we identify and treat these serious conditions.

There are some key differences between the existing SASI model [39] and our model, for instance, we found that the experimental results of SASI are not repeatable. During our experiments, we also found that most of the cases that the model used to flag as uncertain require human intervention. Besides that, the model is unaware that there is a human involved. The selection function in the SASI model is rather ad-hoc because the model's confidence cannot be calibrated so that it accurately reflects the likelihood that the prediction is correct. The set threshold value in the SASI model further reflects that certain ad-hoc choices must be made to make the model perform optimally. The fundamental issue with the SASI model is that it sends all the instances to the human expert which might overwhelm them if the threshold values are not filtered. Our framework is also similar in spirit to the recent CLEFT model [33], but there are several fundamental differences. The key difference is that CLEFT uses a different loss function, and it does not have a greedy mechanism to select the most ideal classifiers. The output from the greedy model maximizes the probability of the model accurately predicting the correct label and refrains from making a prediction when it is uncertain.

Active learning is a machine learning technique that allows algorithms to autonomously select the most informative training data [35]. This can improve the performance of text classification by reducing the need for human annotation and accelerating the learning process. Our work uses active learning by employing a new greedy model exploiting multiple classifiers. Our model only needs to make accurate predictions on a subset of the samples, while the remaining samples are outsourced to a human expert. Our model contributes to the advancement of human-machine collaboration models by focusing on the interaction between the machine and the human expert that will lead to the best joint decision.

The Columbia Suicide Severity Rating Scale (C-SSRS) [32, 39] is a reliable questionnaire used to measure the severity of suicide risk. It has three items: Suicide Ideation, Suicide Behavior, and Suicide Attempt. Each C-SSRS severity class consists of a set of conceptually organized questions that describe the respective category. Responses to these questions across the C-SSRS classes are used to determine an individual's suicide risk [27]. One challenge that researchers face when using the C-SSRS to assess suicide risk from social media content is the wide range of emotions expressed on social media. On social media, non-suicidal users may participate to offer support to others who are deemed suicidal. To address these challenges, [12] defined two additional classes to the existing C-SSRS scale: Suicide Indicator that includes individuals who express suicidal ideation or behaviour on social media, but may not be at immediate risk of suicide. Supportive (Negative class) includes individuals who offer support to others who are suicidal on social media. These two additional classes allow researchers to more accurately assess suicide risk from social media content and to identify individuals who may need support.

3 Our Novel Model (DiGrI)

In this section, we describe the technical details of our novel model, `DiGrI`. Unlike the SASI model [39], `DiGrI` does not use an ad-hoc selection function. Instead, we introduce a principled distorted greedy approach with multiple classifiers that have strong theoretical guarantees [5]. While our optimization model is NP-hard, we propose training an additional model to decide which samples to outsource to a human. This setting differs from the one proposed in [5], which uses only a single maximum-margin classifier. Our classification models are trained on the labelled set. If the model does not make any mistakes on the training set, we can possibly conclude that the samples assigned to the classifier during training are representative of the feature distribution. This means that the classifier will perform well on unseen samples from this distribution. However, we still need to decide whether to outsource an unseen feature vector from the test set to a human expert. This is because outsourcing all unseen samples to humans can be expensive or time-consuming. To address this issue, the model in [5] works by iteratively adding samples to a set of labelled data that will be used to train a classifier.

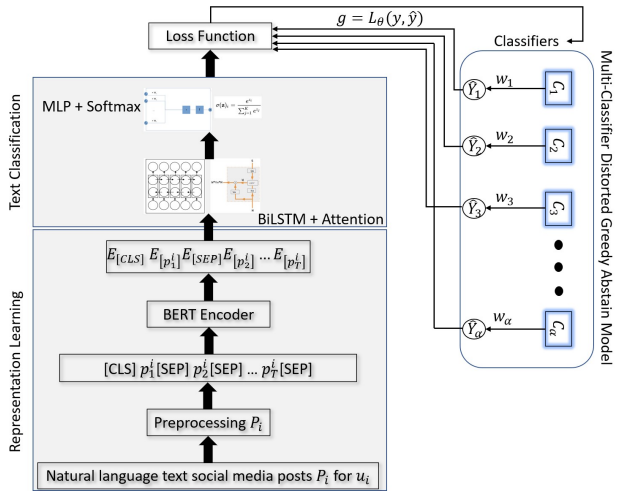

Figure 2: The figure depicting DiGrI for early suicide detection with multiple classifiers exploiting the distorted greedy abstain function. Figures in the text classification component are taken from Wikipedia. [CLS] token represents sentence-level classification and the [SEP] token is used in the model to depict the end of one input and the start of another input in the same sequence input.

At each iteration, the model chooses the sample that is most likely to be misclassified by the classifier and outsources it to a human expert for labelling. The model then updates the classifier based on the newly labelled data. The Distorted Greedy model differs from the standard greedy model in that it distorts the distribution of the labelled data. This is done by giving more weight to samples that are more likely to be misclassified by the classifier.

Problem Formulation: We address the problem of suicide ideation as a classification task. Our goal is to automatically predict the risk of a user committing suicide. To determine the likelihood of suicide for a particular user u_i from a group of users $u_1, u_2, ..., u_{\mathcal{V}}$, where $\mathcal{V}$ denotes the total number of users, we analyzed the post(s) $P_i = p_1^i, p_2^i, \cdots, p_T^i$ where each post is sorted chronologically with T being the final timestamp of u_i of posts that they shared on social media that indicates the most recent post. For the user evaluation, the Columbia-Suicide Severity Rating Scale (C-SSRS) [32] was utilized, following the method in [3]. C-SSRS manually assigns users to five risk levels, from Support (SU) to Attempt (AT), with each level indicating increasing severity. What is defined as a "High" risk can be manually annotated by the expert, e.g., depending upon the user set, one may want every prediction above Behavior to be sent to the human expert as a high-risk instance. This technique has been adopted in [37]. These five levels define our labelled dataset $\mathbf{Y}$. To automatically detect suicide in users, we must also expand the cardinality of the label space by 1, i.e., $|\mathbf{Y}| + 1$ to handle the cases when the model is uncertain in its predictions (meaning that an additional label called "Refrain" is introduced). This label is used when the model is uncertain about the user's suicidal risk.

Overview of DiGrI: Figure 2 shows the architecture of DiGrI. Our model has the following key components: 1) Representation learning model that learns a representation of the input text that captures the relevant features for classification, 2) Text classification model that takes the representation of the input text as input and predicts its class. This model is called the Suicide Ideation Model (SIM) in [39], and 3) Multiple classifier distorted greedy abstain model that decides whether to outsource a sample to a human expert for labelling. While representation learning and classification models, i.e., BiLSTM with attention to text classification are common in the literature [22, 23, 30], our work introduces a novel greedy classifier.

Each social media post made by a user can provide detailed information about the manifestation of suicidal thoughts over time as demonstrated in [28]. To capture this temporal property, we employed the long short-term memory (LSTM) backbones [12, 19, 26, 39]. Each $p_k^i \in P$ is a natural language text post from the dataset. We preprocess the text to make it suitable for the machine learning model. We input the preprocessed text to the vanilla BERT encoder, which outputs a 786-dimensional semantic vector representation for every post's [CLS] token that the user has shared on online social media denoted as $E_k^i = BERT(p_k^i)$. In general, we could exploit any suitable pre-trained language models such as BERT [8] or RoBERTa [24]. However, we used the BERT model in our work due to its strong performance. We then pass these semantic vectors to the text classification model, which is the Bi-LSTM [15] with attention model in our case denoted as $h_k^i = \text{BiLSTM}(E_k^i)$ that gives the hidden states denoted as $\mathbf{x} = [h_1^i, h_2^i, \cdots, h_T^i]$ where $h_k^i \in \mathcal{R}^H$ and H is the latent dimension. The attention model is used to focus on key latent dimensions [39, 41, 50]. We make predictions using the MLP with a softmax layer that gives us $\hat{\mathbf{y}}$, which is a standard setting in several text classification tasks [22, 23, 30].

Our model's abstain function exploits the distorted greedy classifier. This classifier consists of a number, α, of weak learners, trained on the training dataset, that predict the class for each instance with a confidence estimate. In Figure 2, the unique feature of our framework is the greedy model, which consists of a set of classifiers $C_\alpha \in C$ that complement each other in their properties, e.g., discriminative and generative, where $|C|$ is the number of classifiers and $\alpha <= |C|$ and $\forall C_\alpha \in C$, there is an output label $\hat{Y}_\alpha$. In DiGrI, we have $|C| = 10$ due to the computational ease and strong empirical performance. These classifiers make a collective decision whether to abstain or predict the instance confidently. Each classifier predicts an instance with a weight w_α that denotes its confidence in its prediction $\hat{Y}_\alpha$. These weights are jointly trained during the backpropagation phase. At each iteration, the model predicts $\hat{y}$ which is then compared with y through the loss function $L(y, \hat{y})$ parameterised by θ. The Gambler's loss in our model allows the "weighted" gradients, from the classifiers, to propagate through g. The distorted greedy model aggregates the predictions from the trained classifiers and selects one or more trained classifiers for the sample data. The predictions and their confidence are then sent to human experts.

Our objective function can be expressed as the difference between two functions, $f = g - \vartheta$, where g is monotone, non-negative, and γ-weakly submodular, and ϑ is non-negative and modular as demonstrated in [5] for a single classification function. The parameter γ represents the degree to which g is more sensitive to the marginal contribution of recently added data points. ϑ is a non-negative and modular function that accounts for the cost of outsourcing a

data point to humans. It represents a fixed cost associated with each data point outsourced, regardless of the order in which they are chosen. This allows us to exploit a recently introduced deterministic greedy model [5], as well as a more efficient randomized variant of the model, to obtain approximation guarantees for solving the problem. This model aims to find a subset of training data points to outsource to humans for labelling while still maintaining high overall classification accuracy. The distorted greedy model is based on the concept of γ-weakly submodular functions. These functions are a generalization of submodular functions, which are known to have good approximation guarantees for certain optimization problems. γ-weakly submodular functions capture the benefit of outsourcing a data point to humans.

The distorted greedy model in our model with multiple classifiers works by iteratively selecting the data point that maximizes its marginal contribution to the difference between g and ϑ. This means that at each step, the model chooses the data point that will most improve the overall classification accuracy while also minimizing the total cost of outsourcing data points. The model can find a solution that is within a constant approximation factor of the optimal solution, making it a valuable tool for optimizing classification models under human assistance. This model is jointly optimized with a loss function called the Gambler's loss, which is the same loss function used in [25, 39]. This joint learning ensures that the parameters of the model are faithfully trained in a single coherent space, resulting in improved performance without the need for a threshold parameter. Overall, this model provides a principled approach to optimizing classification models under human assistance, ensuring that the overall accuracy remains high while minimizing the reliance on human labelling.

We denote $\mathcal{S}$ as the subset of the samples sent to humans where $\mathcal{S} \subset \mathcal{V}$ and $|\mathcal{S}| \leq \beta$ where β is the small subset of users. To incorporate the cost of outsourcing data, we introduce, into the loss function, a component that is proportional to the number of samples outsourced to humans. This term encourages the classifier to make accurate predictions on the data that it is trained on so that it does not need to rely on human experts as much. The loss function is also proportional to the uncertainty of the classifier's predictions. As a result, we encourage the classifier to outsource samples that it is unsure about so that it can improve its performance on these samples. The loss function, $L_\theta(y, \hat{y})$ in our model is denoted by:

$$\min_{\mathcal{S},\theta} \sum_{j=1}^{\alpha} \sum_{i \in \mathcal{V} \setminus \mathcal{S}} l_j(h_\theta(\mathbf{x}_i, y_i)) + \sum_{i \in \mathcal{S}} \vartheta(\mathbf{x_i}, y_i), \text{ s.t. } |\mathcal{S}| \leq \beta \quad (1)$$

4 Experiments and Results

4.1 Experimental Settings

Dataset: 2.0 Experimental Advanced. Lacks access to real-time info and some Gemini features. To assess `DiGrI`'s performance, a public dataset from [12], focused on identifying suicide risk on Reddit, was employed. This dataset comprises 500 users, each categorized into one of five escalating risk levels based on their activity across nine subreddits related to mental health and suicide. Four practicing psychiatrists labeled the data utilizing C-SSRS guidelines. Table 1 summarizes the distribution of these risk levels (class labels). This dataset has been widely used to evaluate computational models for suicide ideation.

Table 1: Dataset statistics

Label	Percentage (%)
Supportive (SU)	20%
Suicidal Indicator (IN)	20%
Suicidal Ideation (ID)	34%
Suicidal Behavior (BR)	15%
Actual Attempt (AT)	9%

While several datasets, including those presented in Ghanadian et al. [13], focus on suicide ideation, we argue that the dataset introduced by Gaur et al. [12] is most suitable for our research. This is primarily due to two key factors: 1) the explicit inclusion of risk levels, providing a granular understanding of suicide ideation, and 2) its alignment with other Reddit-based datasets, such as the UMDs Reddit Suicidality Dataset [13], facilitating comparative analysis.

Preprocessing: To mitigate noise and address out-of-vocabulary (OOV) terms, our pre-processing pipeline includes several steps. Initially, spelling errors were rectified, and emoticons/emojis were replaced with corresponding textual representations. Hashtag symbols (#) were removed to separate conjoined words. URLs, numbers, user mentions, contractions, and lengthened words were standardized. These steps were performed using the emoji and ekphrasis Python libraries. Additionally, punctuation, repeated words, and stopwords were removed using standard procedures and regular expressions.

Hyperparameters: To maintain consistency, experiments were conducted on the training set utilizing 5-fold cross-validation, with 80 users per fold, consistent with previous research [12, 36]. Hyperparameter tuning was performed using grid search. The number of layers (n) was optimized by testing values within the set (n) $\in \{1, 2, 3\}$. Other hyperparameters were also varied: dropout rate δ from 0 to 0.8 in increments of 0.2, hidden dimension (H) from 32 to 128 in increments of 32, learning rate (lr) from 0.001 to 0.01 in increments of 0.004, and the regularizer parameter β from 0 to 3.0 in increments of 0.3. Optimizers used included Adam, Adamax, and AdamW, with a batch size of 16. Varying post lengths were addressed by padding during training, and the model was trained for 150 epochs. Optimal hyperparameters were determined as: (n) = 2, lr = 0.005, δ = 0.5, O = AdamW, and (H) = 128.

Evaluation metrics: To evaluate the model's performance on the coverage samples, we use graded variants of F1 score, Precision, and Recall, as described by [12]. Following the work of [12, 39], we use the following metrics:

$$FN = \frac{\sum_{i=1}^{N} I(k_i^a > k_i^p)}{N} \quad FP = \frac{\sum_{i=1}^{N} I(k_i^p > k_i^a)}{N} \quad (2)$$

In the equations above, in [12], the authors modified the definitions of false positives (FP) and false negatives (FN). FN is defined as the ratio of the number of times the predicted suicide risk level k^p is less than the actual risk level k^a to the total number of samples N. FP is defined as the ratio of the number of times the predicted risk k^p is greater than the actual risk k^a to N.

Let P_T represent the total number of test samples, $P_{corr+refrain}$ denote the count of samples either accurately predicted or abstained

from, $P_{refrain}$ indicate the total number of abstained samples, and P_{in} signify the number of incorrect predictions within the abstained group. We further define two metrics: Robustness and Fail-Safe Rejects (FSR).

$$Robustness = \frac{P_{corr+refrain}}{P_T}, FSR = \frac{P_{in}}{P_{refrain}} \quad (3)$$

Robustness measures the proportion of samples that are either correctly categorized or flagged for immediate review. Fail-safe rejects quantify the percentage of rejected samples that were, in fact, incorrect. A higher fail-safe rejects score indicates reduced workload for human moderators, as fewer non-critical samples will require their attention.

4.2 Baseline Models

For use in the traditional models described below, we concatenated several language-based features (LBFs) into a single vector for each post. These LBFs comprised the psychological Linguistic Inquiry and Word Count (LIWC), part-of-speech counts, and term frequency-inverse document frequency [34].

Traditional Methods: We tested four traditional machine learning models: 1) **SVM+RBF** [1]: This classifier employs an SVM with a Radial Basis Function kernel and hinge loss as the objective function, using the previously generated LBF vector as input. 2) **SVM-L** [1]: This model utilizes an SVM with a linear kernel ($c = 1.5$) and hinge loss objective function, taking the aforementioned LBF vector as input. 3) **RF** [38]: This approach uses a Random Forest classifier with the Gini Impurity metric as the objective function, utilizing the previously derived LBFs. 4) **MLP** [1]: employs a Multilayer Perceptron with two hidden layers, each containing 64 neurons. It uses the LBFs as input and log loss as the objective function.

Deep Learning Methods: We evaluated seven deep learning models: 1) **Context CNN** [12]: This model utilizes GloVe embeddings [31] to represent user posts, which are then concatenated and processed by a contextual CNN. 2) **Suicide Detection Model (SDM)** [3]: Posts are converted into fine-tuned FastText embeddings and subsequently fed into an attention-based long-short-term memory (LSTM) network. 3) **ContextBERT** [26]: The winning model of the 2019 CL Psych Classification Competition [49], this model encodes Reddit posts using BERT and passes them to a GRU. 4) **SISMO** [36]: Word representations generated using Longformer are passed to an attention-based bidirectional LSTM (BiLSTM) with ordinal loss as the objective function. 5) **MentalBERT:** [20]: A recently developed transformer-based model, a variant of BERT, specifically designed for text mining in mental health. 6) **MentalLLaMa:** [47]: A recent model based on LLaMa [44] that formulates interpretable mental health analysis as a text generation task. 7) **SASI** [39]: A risk-averse, self-aware transformer-based hierarchical attention classifier designed to abstain from uncertain predictions.

4.3 Experimental Results

Overall comparison: The results in Table 2 demonstrate the superior performance of our proposed method compared to several baseline approaches in identifying online health information related to suicide risk. Deep learning models, including DiGrI, surpass traditional methods like SVM with RBF, SVM-L, RF, and MLP, which depend on manually engineered features. This improved performance is due to deep learning models' ability to effectively capture contextual information and the complexities of a user's mental state.

Table 2: DiGrI was evaluated against multiple baselines, demonstrating a statistically significant improvement ($p < 0.005$) over the second-best performing method (underlined), as determined by the Mann-Whitney U test.

Model	GP	GR	FScore	FSR	Robustness
SVM+RBF	0.53	0.51	0.52	-	-
SVM-L	0.60	0.45	0.52	-	-
RF	0.68	0.49	0.57	-	-
MLP	0.45	0.59	0.51	-	-
Contextual CNN	0.65	0.52	0.59	-	-
ContextBERT	0.61	0.57	0.60	-	-
SISMO	0.62	0.62	0.62	-	-
MentalBERT	0.64	0.64	0.64	-	-
MentaLLaMA-7B	0.66	0.65	0.65	-	-
SASI (Cov 100%)	0.65	0.54	0.60	-	0.43
SASI (Cov 85%)	0.65	0.55	0.59	0.75	0.51
SASI (Cov 50%)	0.65	0.55	0.61	0.58	0.81
DiGrI (Cov 100%)	0.68	0.62	0.65	**1**	0.25
DiGrI (Cov 85%)	0.72	0.65	0.69	**1**	0.36
DiGrI (Cov 50%)	**0.77**	**0.85**	**0.81**	0.96	**0.76**

Sequential deep learning-based methods, such as DiGrI, showcase superior performance compared to CNN-based methods, possibly due to their capacity to capture more extended temporal context. DiGrI consistently outperforms SASI across different coverage levels, signifying its robustness in handling varying degrees of content coverage. Most importantly, we also outperform recently developed highly effective language models MentalBERT and MentalLlaMa.

Furthermore, our findings underscore the significance of DiGrI's ability to avoid committing to erroneous predictions, as indicated by its superior FSR (Fail-Safe Reject) scores. The Mann-Whitney U test confirms the statistical significance ($p < 0.005$) of these performance improvements, emphasizing the superiority of DiGrI over the tested baseline methods. We conclude that DiGrI stands out as an effective and robust approach for identifying suicide risk in online health information. Its superior performance across various coverage levels and the statistical significance of its improvements reinforce its potential as a reliable tool for content moderation in mental health contexts.

Coverage and Performance Trade-off: We conducted an evaluation of the proposed Greedy classifier and SASI across a spectrum of target coverage values, adjusting the threshold parameter in SASI. As illustrated in Figure 3 and supported by the data in Table 2, the trade-off between coverage and performance becomes apparent.

For the DiGrI, maintaining a high Fail-Safe Reject (FSR) of 1.00 at all coverage levels, it achieves a competitive FScore, with values ranging from 0.61 to 0.99. This emphasizes the DiGrI's strength in consistently avoiding false negatives, particularly evident in its FScore performance at 100% coverage. In contrast, SASI demonstrates varying performance metrics across coverage levels. While achieving a respectable FScore, its FSR values fluctuate, suggesting a trade-off between false negatives and false positives. Notably, DiGrI

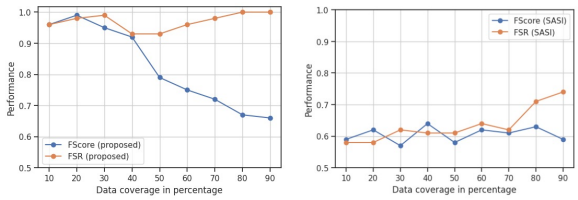

Figure 3: DiGrI (left) v/s SASI (right): changes in performance metrics with increasing coverage

consistently outperforms SASI in terms of FScore, showcasing the effectiveness of our approach. Our analysis reveals that DiGrI strikes a favourable balance between coverage and performance. At 90% coverage, it outperforms SASI statistically, maintaining a perfect FSR score of 1.00. This underscores the potential of DiGrI in achieving competitive performance while efficiently moderating content. We conclude that DiGrI exhibits adaptability to varying coverage requirements, offering competitive performance compared to the state-of-the-art SASI model. The results underscore the importance of careful consideration in selecting the optimal coverage threshold for real-world deployment, with DiGrI emerging as a promising solution in achieving a balanced trade-off between coverage and performance.

Taking into account the performance trade-off, we propose that the optimal data coverage for DiGrI lies between 50% and 60%. This range allows for robust model performance while ensuring a manageable workload for human moderators. This underscores the potential of DiGrI in balancing performance and moderation workload, even in the face of trade-offs.

Ablation analysis: We systematically investigate the impact of introducing individual components to our proposed model under varying coverage settings (Table 3). Initially, employing the Suicide Ideation Model (SIM) [37] as the base model yields a consistent FScore of 0.53 across different coverage rates (50%, 85%, and 100%). Subsequently, augmenting SIM with the Gambler Loss (GL) [37] demonstrates a notable enhancement, resulting in FScore improvements to 0.61, 0.59, and 0.60 for 50%, 85%, and 100% coverage, respectively. Remarkably, the most substantial performance gains are observed in DiGrI. The FScore experiences a significant boost, reaching 0.81, 0.69, and 0.61 for the respective coverage rates. This underscores the strength of DiGrI in improving the model's predictive capabilities, particularly evident in scenarios where full coverage is essential. In conclusion, the cumulative effect of the combined components, especially the distorted Greedy model, results in a robust and high-performing model across diverse coverage settings, affirming its efficacy in enhancing the overall predictive accuracy of the proposed approach.

Analysing different classifiers: The evaluation of Mean Squared Error (MSE) across varying coverage rates reveals nuanced performance patterns among different classifiers (Table 4). The distorted Greedy classifier consistently demonstrates superior predictive accuracy, yielding lower MSE across the entire spectrum of coverage rates. This consistent performance underscores the reliability of the distorted Greedy classifier in minimizing errors. Other classifiers exhibit more variable behaviour, with fluctuations and increasing

Table 3: Ablation analysis: F-scores are averaged across 10 folds. The asterisk (*) denotes that the proposed method yielded a significant performance improvement ($p < 0.05$) compared to other variants, as determined by the Mann-Whitney U test.

Model	Cov (50%)	Cov (85%)	Cov (100%)
SIM	0.53	0.53	0.53
SIM+Gambler loss (GL)	0.61	0.59	0.60
DiGrI	0.81*	0.69*	0.61*

Table 4: MSE Graphs

Classifier\Cov	0.2	0.4	0.6	0.8	1
AdaBoost	0.45	0.56	0.68	1.22	1.12
MLP	0.20	0.65	0.47	1.29	1.13
KNeighbor	0.36	0.32	0.40	0.93	1.25
RandomForest	0.23	0.30	0.97	1.10	1.04
Gaussian Process	0.86	1.31	0.80	1.02	1.31
Decision Tree	0.38	0.36	1.33	0.94	0.82
XGB	1.47	0.83	0.97	1.32	1.29
GaussianNB	0.96	0.50	1.57	1.54	1.29
Gradient Boosting	0.18	0.30	0.47	0.95	1.00
Distorted Greedy	0.08	0.10	0.33	0.78	0.88

trends in MSE across different coverage rates. The findings highlight the robustness of the distorted Greedy classifier and emphasize its potential suitability for applications demanding consistent and accurate predictions. However, it is also crucial to consider model interpretability and computational efficiency when selecting the most appropriate classifier for specific tasks.

The analysis of various classifiers with FScore across diverse coverage rates reveals distinctive performance patterns (Table 5). AdaBoost, MLP, KNeighbor, and RandomForest exhibit fluctuating performances, displaying variable FScore values across different coverage rates. Gaussian Process and Decision Tree also demonstrate diverse performances. XGB and Gradient Boosting exhibit variability in FScore, while GaussianNB demonstrates fluctuations peaking at higher coverage rates. Notably, the Greedy model consistently outperforms other classifiers, showcasing an ascending trend in FScore values with increasing coverage rates. This steadfast performance of the Greedy model underlines its efficacy in achieving higher predictive accuracy, making it the optimal choice for suicide risk assessment within this analytical framework.

Qualitative Analysis: The key advantage of DiGrI is in its ability to exercise caution and refrain from making misleading predictions, particularly over high-risk samples. In our study involving 4 users (Figure 4), we highlight instances where DiGrI demonstrates its nuanced decision-making. Notably, DiGrI refrains from committing to predictions for high-risk users, assigning them a high priority for immediate review and response. Even when DiGrI correctly predicts the risk level of user C, it chooses to refrain, possibly due to a cautious approach prompted by phrases such as "take my life" in the user's timeline. This cautious prioritization is indicative of

Table 5: Difference classifiers v/s coverage.

Coverage	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
AdaBoost	0.45	0.56	1.10	0.68	1.12	1.22	0.74	0.88
MLP	0.20	0.65	0.68	0.47	0.94	1.29	2.91	1.13
KNeighbor	0.36	0.32	0.42	0.30	0.23	0.93	0.89	1.25
RandomForest	0.23	0.20	0.21	0.97	0.77	1.10	0.74	1.04
Gaussian Process	0.86	1.31	0.88	0.80	1.11	1.02	1.39	1.31
Decision Tree	0.38	0.36	0.08	1.33	1.55	0.94	1.19	0.82
XGB	1.47	0.83	1.84	0.97	1.34	1.32	1.28	1.29
GaussianNB	0.96	0.50	1.52	1.57	1.34	1.54	1.74	1.29
Gradient Boosting	0.08	0.30	0.12	0.33	0.51	0.95	1.28	1.00
Distorted Greedy	0.08	0.10	0.12	0.47	0.54	0.78	0.91	1.13

DiGrI's commitment to ensuring the utmost accuracy, especially for users already at a relatively high risk.

Examining User B, who exhibits a very low sign of risk, DiGrI confidently highlights without the need to refrain. Conversely, for User D, DiGrI makes a confident yet erroneous prediction. Despite the misstep, as the user is not deemed high risk, DiGrI assigns the same priority level as the true risk label. While this particular example may not be a cause for concern, it sheds light on scenarios where DiGrI, despite confidence, may assign a low-risk score to a high-risk user.

The qualitative analysis presented in Figure 4 illuminates insights into the predictive performance of DiGrI relative to the strong baseline, SASI. The visual representation in Figure 4 provides a nuanced examination of ground truth instances, revealing the heightened decision-making acumen of DiGrI. Particularly noteworthy is the case of User C's post, wherein DiGrI demonstrates a judicious approach by refraining from making a prediction, thereby mitigating the risk of potential inaccuracies observed in the original SASI code. This discerning behaviour underscores the efficacy of the embedded distorted Greedy model within DiGrI, attesting to its capability to navigate uncertainty judiciously.

Furthermore, the instances involving Users A, 198, and 332 underscore the consistent superiority of DiGrI over the original SASI code. DiGrI exhibits a commendable ability to accurately predict outcomes where SASI encounters limitations, further corroborating its enhanced predictive accuracy. The qualitative findings collectively emphasize the practical advantages of DiGrI in content moderation applications, showcasing its propensity to not only outperform the baseline but also exercise caution and refrain from predictions in instances where SASI falters. These observations contribute valuable insights into the nuanced decision-making capabilities of DiGrI and its potential as an advanced tool in the domain of content moderation.

4.4 Discussion

By leveraging trained classifiers, DiGrI effectively resolves the issue of inconsistent FSR scores. These classifiers, adept at distinguishing between refrain and non-refrain samples, contribute to the overall robustness. The distorted Greedy abstain function, guided by the classifier with the highest refrain accuracy, ensures consistently high FSR across diverse coverage values (10 - 100%). DiGrI offers

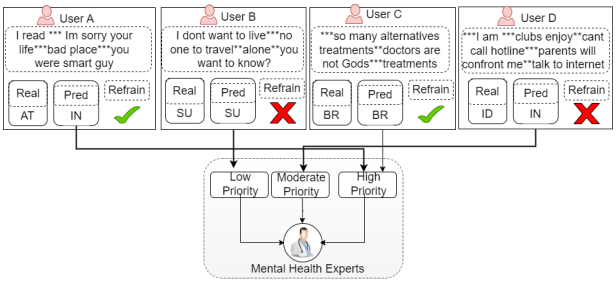

Figure 4: DiGrI leverages user prioritization. We display actual user labels alongside predicted labels, also noting when DiGrI opted to abstain from prediction. Additionally, we illustrate how DiGrI categorizes users into priority levels. To safeguard user privacy, all examples have been paraphrased using a moderate disguise strategy.

a superior and precise solution to challenges encompassing low FScore due to diminished prediction accuracy, uncertainty in refrain weight distribution from the abstain function, low robustness score resulting from inaccurate refrains, and sensitivity issues related to adjusting coverage values.

Inconsistent Fail-Safe Rejects: The SASI model [39] exhibited inconsistent Fail-Safe Rejects (FSR) scores. In an attempt to rectify this issue, the authors experimented with various loss functions, such as Gambler's loss, Ordinal loss, Gambler plus Ordinal loss, and Cross Entropy loss. Notably, the highest graded metric scores were achieved with the Gambler's loss function. However, it was observed that Gambler's loss function only led to an increase in FSR with data coverage higher than 70%.

In contrast, DiGrI employs trained classifiers for the classification of refrain and non-refrain samples. The heightened accuracy of these classifiers makes them well-suited for discerning the samples warranting refrain. By allowing the distorted Greedy abstain function to select the trained classifier with the highest refrain accuracy, DiGrI consistently achieves high FSR across coverage values ranging from 10% to 100%. This strategic approach enhances the robustness of DiGrI, providing a more reliable solution to the challenge of inconsistent FSR scores, as demonstrated in the results.

5 Conclusions

We introduce a new model for suicide detection that integrates selective prioritization into deep learning-based risk assessment. DiGrI exhibits intrinsic self-awareness, choosing to refrain from predictions under uncertainty and instead designating these instances for high-priority review. Quantitative evaluation on real-world data confirms DiGrI's effectiveness, demonstrating that it successfully avoided high-risk situations by abstaining from incorrect predictions in 83% of cases.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Granted No. 62276190).

References

[1] Payam Amini, Hasan Ahmadinia, Jalal Poorolajal, and Mohammad Moqaddasi Amiri. 2016. Evaluating the high risk groups for suicide: a comparison of logistic regression, support vector machine, decision tree and artificial neural network. *Iranian journal of public health* 45, 9 (2016), 1179.

[2] Vahid Balazadeh Meresht, Abir De, Adish Singla, and Manuel Gomez Rodriguez. 2022. Learning to switch among agents in a team. *Transactions on Machine Learning Research* 2022, 7 (2022), 1–30.

[3] 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)*. 1718–1728.

[4] Caroline Coope, David Gunnell, William Hollingworth, Keith Hawton, Nav Kapur, Vanessa Fearn, Claudia Wells, and Chris Metcalfe. 2014. Suicide and the 2008 economic recession: who is most at risk? Trends in suicide rates in England and Wales 2001–2011. *Social Science & Medicine* 117 (2014), 76–85.

[5] Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez Rodriguez. 2021. Classification under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 5905–5913.

[6] Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology* (2023), 1–14.

[7] SP Devika, MR Pooja, MS Arpitha, and Ravi Vinayakumar. 2023. BERT-Based Approach for Suicide and Depression Identification. In *Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022*. Springer, 435–444.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[9] Fahim Faisal, Mirza Muntasir Nishat, Kazi Raine Raihan, Ahmad Shafiullah, and Sanjida Ali. 2023. A Machine Learning Approach for Analyzing and Predicting Suicidal Thoughts and Behaviors. In *2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 43–48.

[10] Yevgeniy Feyman, Stuart M Figueroa, Yingzhe Yuan, Megan E Price, Aigerim Kabdiyeva, Jonathan R Nebeker, Merry C Ward, Paul R Shafer, Steven D Pizer, and Kiersten L Strombotne. 2023. Effect of mental health staffing inputs on suicide-related events. *Health services research* 58, 2 (2023), 375–382.

[11] Neda Firoz, Olga Grigorievna Beresteneva, Aksyonov Sergey Vladimirovich, Mohammad Sadman Tahsin, and Faiza Tafannum. 2023. Automated Text-based Depression Detection using Hybrid ConvLSTM and Bi-LSTM Model. In *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*. IEEE, 734–740.

[12] 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*. 514–525.

[13] Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2024. Socially Aware Synthetic Data Generation for Suicidal Ideation Detection Using Large Language Models. *IEEE Access* (2024).

[14] Noman Ghiasi, Yusra Azhar, and Jasbir Singh. 2023. Psychiatric illness and criminality. In *StatPearls [internet]*. StatPearls Publishing.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780.

[16] Jennifer A Hoffmann, Megan M Attridge, Michael S Carroll, Norma-Jean E Simon, Andrew F Beck, and Elizabeth R Alpern. 2023. Association of youth suicides and county-level mental health professional shortage areas in the US. *JAMA pediatrics* 177, 1 (2023), 71–80.

[17] Tarek Ibrahim, Amr Gebril, Mohammed K Nasr, Abdul Samad, Hany A Zaki, and Mohammed Nasr Sr. 2023. Exploring the Mental Health Challenges of Emergency Medicine and Critical Care Professionals: A Comprehensive Review and Meta-Analysis. *Cureus* 15, 7 (2023).

[18] Loukas Ilias, Spiros Mouzakitis, and Dimitris Askounis. 2023. Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media. *IEEE Transactions on Computational Social Systems* (2023).

[19] Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2022. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications* 34, 13 (2022), 10309–10319.

[20] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 7184–7190. https://aclanthology.org/2022.lrec-1.778

[21] Fazida Karim, Azeezat A Oyewande, Lamis F Abdalla, Reem Chaudhry Ehsanullah, and Safeera Khan. 2020. Social media use and its connection to mental health: a systematic review. *Cureus* 12, 6 (2020).

[22] Chenbin Li, Guohua Zhan, and Zhihua Li. 2018. News text classification based on improved Bi-LSTM-CNN. In *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*. IEEE, 890–893.

[23] Gang Liu and Jiabao Guo. 2019. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* 337 (2019), 325–338.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[25] 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. *Advances in Neural Information Processing Systems* 32 (2019).

[26] 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*. 39–44.

[27] William V McCall, Ben Porter, Ashley R Pate, Courtney J Bolstad, Christopher W Drapeau, Andrew D Krystal, Ruth M Benca, Meredith E Rumble, and Michael R Nadorff. 2021. Examining suicide assessment measures for research use: Using item response theory to optimize psychometric assessment for research on suicidal ideation in major depressive disorder. *Suicide and Life-Threatening Behavior* 51, 6 (2021), 1086–1094.

[28] John L Oliffe, John S Ogrodniczuk, Joan L Bottorff, Joy L Johnson, and Kristy Hoyak. 2012. "You feel like you can't live anymore": Suicide from the perspectives of Canadian men who experience depression. *Social science & medicine* 74, 4 (2012), 506–514.

[29] Urvashi Panchal, Gonzalo Salazar de Pablo, Macarena Franco, Carmen Moreno, Mara Parellada, Celso Arango, and Paolo Fusar-Poli. 2023. The impact of COVID-19 lockdown on child and adolescent mental health: systematic review. *European child & adolescent psychiatry* 32, 7 (2023), 1151–1177.

[30] MR Pavan Kumar and Prabhu Jayagopal. 2023. Context-sensitive lexicon for imbalanced text sentiment classification using bidirectional LSTM. *Journal of Intelligent Manufacturing* 34, 5 (2023), 2123–2132.

[31] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[32] Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, et al. 2011. The Columbia–Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American journal of psychiatry* 168, 12 (2011), 1266–1277.

[33] Sujit Roy, Vishal Gaur, Haider Raza, and Shoaib Jameel. 2023. CLEFT: Contextualised Unified Learning of User Engagement in Video Lectures With Feedback. *IEEE Access* 11 (2023), 17707–17720.

[34] Gerard Salton and Michael J McGill. 1983. *Introduction to modern information retrieval*. mcgraw-hill.

[35] Annie Sauer, Robert B Gramacy, and David Higdon. 2023. Active learning for deep Gaussian process surrogates. *Technometrics* 65, 1 (2023), 4–18.

[36] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021. Towards ordinal suicide ideation detection on social media. In *Proceedings of the 14th ACM international conference on web search and data mining*. 22–30.

[37] Ramit Sawhney, Harshit Joshi, Rajiv Ratn Shah, and Lucie Flek. 2021. Suicide Ideation Detection via Social and Temporal User Representations using Hyperbolic Learning. In *NAACL-HLT*. Association for Computational Linguistics, 2176–2190.

[38] 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*. 91–98.

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

[40] Soomin Shin and Kyungwon Kim. 2023. Prediction of suicidal ideation in children and adolescents using machine learning and deep learning algorithm: a case study in South Korea where suicide is the leading cause of death. *Asian journal of psychiatry* 88 (2023), 103725.

[41] 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*. 8124–8137.

[42] Surendrabikram Thapa, Mohammad Salman, Siddhant Bikram Shah, Shuvam Shiwakoti, Qi Zhang, Liang Hu, Imran Razzak, and Usman Naseem. 2024. THYMES: A Framework for Detecting Suicidal Ideation from Social Media Posts Using Hyperbolic Learning. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 6538–6546.

[43] Surendrabikram Thapa, Mohammad Salman, Siddhant Bikram Shah, Qi Zhang, Junaid Rashid, Liang Hu, Imran Razzak, and Usman Naseem. 2024. SAFENet: Towards a Robust Suicide Assessment in Social Media Using Selective Prediction Framework. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 660–669.

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[45] Osman Ulvi, Ajlina Karamehic-Muratovic, Mahdi Baghbanzadeh, Ateka Bashir, Jacob Smith, and Ubydul Haque. 2022. Social media use and mental health: A global analysis. *Epidemiologia* 3, 1 (2022), 11–25.

[46] Michelle M Vance, Jeannette M Wade, Mervin Brandy Jr, and Aiyana Rice Webster. 2023. Contextualizing Black women's mental health in the twenty-first century: Gendered racism and suicide-related behavior. *Journal of racial and ethnic health disparities* 10, 1 (2023), 83–92.

[47] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. In *Proceedings of the ACM on Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 4489–4500. https://doi.org/10.1145/3589334.3648137

[48] Jianlong Zhou, Hamad Zogan, Shuiqiao Yang, Shoaib Jameel, Guandong Xu, and Fang Chen. 2021. Detecting community depression dynamics due to covid-19 pandemic in australia. *IEEE Transactions on Computational Social Systems* 8, 4 (2021), 982–991.

[49] Ayah Zirikly, Philip Resnik, Ozlem 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*. 24–33.

[50] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. Depressionnet: learning multi-modalities with user post summarization for depression detection on social media. In *proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 133–142.