2024 IEEE International Conference on Big Data (BigData) | 979-8-3503-6248-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigData62323.2024.10825048

Overview of IEEE BigData 2024 Cup Challenges: Suicide Ideation Detection on Social Media

Jun Li
The Hong Kong Polytechnic University
Hong Kong, China
hialex.li@connect.polyu.hk

Yifei Yan
City University of Hong Kong
Hong Kong, China
yfyan8-c@my.cityu.edu.hk

Ziyan Zhang
The Hong Kong Polytechnic University
Hong Kong, China
ariana.zhang@connect.polyu.hk

Xiangmeng Wang
The Hong Kong Polytechnic University
Hong Kong, China
xiangmengpoly.wang@polyu.edu.hk

Hong Va Leong
The Hong Kong Polytechnic University
Hong Kong, China
cshleong@comp.polyu.edu.hk

Nancy Xiaonan Yu
City University of Hong Kong
Hong Kong, China
nancy.yu@cityu.edu.hk

Qing Li*
The Hong Kong Polytechnic University
Hong Kong, China
qing-prof.li@polyu.edu.hk

***Abstract*—This overview presents one of the cup challenges of IEEE BigData 2024, with the topic of suicide risk level detection on social media posts. Given a training set of N = 2000 posts (N = 500 labelled and N = 1500 unlabelled posts) from r/SuicideWatch subreddits, the task of this challenge is to develop a predictive model capable of classifying the suicidal posts into four levels (i.e., indicator, ideation, behaviour, and attempt). The dataset provided simulated the obstacles existed in relevant fields (e.g., model overfitting, data scarcity and class imbalance), participating teams are supposed to tackle these issues while exploring the effectiveness of various model architectures. We received submissions from 21 teams and works of 13 teams underwent final evaluation. Teams addressed key challenges in suicide risk detection including limited suicidal data and suicidal risk imbalance. They employed novel approaches to overcome these obstacles, leveraging a diverse range of models from foundational base language models (BLMs) to state-of-the-art large language models (LLMs). In the competition, the highest weighted F1-score achieved under the final evaluation was 0.7605. The findings of this challenge can provide technical implications to social media suicide detection and contribute the clinical effectiveness to the applications of machine learning in digital suicide or mental healthcare management.**

***Index Terms*—Data Mining Competitions, Suicide Risk, Suicide Detection, Social Media, Big Data Processing, Overview.**

I. Introduction

Suicide remains a serious global public health concern. The latest statistics showed that the global crude suicide rate was 9.2 per 100 000 population, making it one of the leading causes of death worldwide[1]. Suicide is preventable with timely detection and regular monitoring being the key steps [1]. A landmark meta-analysis of suicide research in the past 50 years suggested that dominating prediction works (e.g., biological, sociological, and psychological approaches) showed limited performance and minimal improvements over time [2]. Advances in machine learning (ML) enabled suicide prediction to achieve better accuracy over traditional methods [3]–[5]. Despite the superior performance, ML algorithms continue to fall short of clinical relevance due to issues such as inflated prediction estimation, low out-sample prediction validity, and failure to address imbalanced outcome [6]–[9]. This competition underscores the needs of improvements over existing ML algorithms in suicide prediction and calls for creative models with greater clinical effectiveness.

Due to its flexibility, automation, and capability to deal with complex dataset, machine learning is assumed to revolutionize the overall healthcare systems [10], [11]. One application of ML models in suicide detection involves the natural language processing task of classifying suicide risk levels among online social media posts data (e.g., posts from Twitter, Reddit, and Weibo). Popular models designed for text-based suicide classifications ranging from traditional methods such as support vector machine, decision trees, ensemble tree [12]–[14], to deep learning methods such as convolutional neural network, bidirectional encoder representations from transformers, long short-term memory network, and attention mechanism [15]–[18], and hybrid/ensemble approach [19]–[22]. These ML models were found to achieve great performance, with accuracy ranging from 73.6% to 98.5% [23]. However, these BLMs bear the problem of poor generalisability to new data [24]. More effective models are needed for the accurate suicide detection.

The last two years witnessed the burgeoning of LLMs and studies integrated LLMs into suicide detection are dominating the field. Given instructions, these LLMs (e.g., GPT-3.5, GPT-4, LLaMA, and Alpaca) are asked to infer the suicide risk of target posts, with or without reasons required, in zero-, few-shot, and finetune settings. Despite the initial potentiality,

* Corresponding author.

[1] https://www.who.int/data/gho/data/themes/mental-health/suicide-rates

979-8-3503-6248-0/24/$31.00 ©2024 IEEE

their performance is limited and unsatisfactory. For example, the use of GPT-3.5 to classify suicide risk result only an F1 score of 0.37 and the performance was not superior than the state-of-the-art NLP models in zero or few-shot manners [25], [26]. Existing LLMs are basically general-purpose without being trained by suicide data, and domain-specific fine-tuning can be an effective solution to improve their accuracy [27]. After instruction-finetuning using suicide dataset, the best-finetuned LLM models (i.e., Mental-FLAN-T5 and MentaLLaMA) showed better performance than top BLMs (e.g., 86.6% vs 85.3% in accuracy and 75.58 vs 72.16 in weighted F1-score), and a maximum of 25.1% increase in accuracy compared with non-finetuned models [24], [28]. Additionally, the latest update of the CLPsych 2024 Shared Task concluded the prevailing use of combing instruction-finetuned LLMs with traditional NLP among participants [29]. The mean precision/recall rate of such approach reached above 90% in determining suicide risk levels and extracting supporting post content [29]. The flexible and complimentary integration of different approaches seems to be promising. Nevertheless, its apparent excellence may be a pseudomorph resulting from overfitting, as some of these advanced models have not been subjected to cross-sample validation [8], [28].

The performance of ML models is predominantly determined by the quantity and quality of the training data utilized. Available dataset of suicide-relevant social media posts with labeled risk level was scarce. Shing et al. [30] created the University of Maryland Reddit Suicidality Dataset (Version 2) containing 621 users from r/SuicideWatch subreddits and assessed user-level suicide risk (i.e., No, Low, Moderate, or Severe) based on their published posts in r/SuicideWatch subreddits. The risk level of another 500 Redditors were annotated by [31] using the 5-label Columbia Suicide Severity Rating Scale (i.e., supportive, indicator, ideation, behavior, and attempt) based on posts from other subreddits except r/SuicideWatch. The more fine-grained annotation of both post-level and context-aware user-level suicide risk was conducted by [32], where 500 users along with their 3998 r/SuicideWatch posts were classified into indicator, ideation, behavior, and attempt level. Datasets containing posts from other social media platforms and languages were also recorded, despite some of them are unavailable [14], [33], [34]. Moreover, existing datasets faced the issue of uneven risk label distributions, with fewer samples labeled for higher suicide risk (e.g., behavior, attempt, and death). For example, the proportions of ideation labels were substantially higher than the labels of attempt level in both [32] (39% vs 9.2%) and [31] (34% vs 9%). Such issue corresponded with the facts that lower suicide risk were usually more prevalent than suicide high-risk states, revealing the progressive nature of the suicide risk development [35], [36]. However, these imbalanced datasets can result in biased and erroneous predictions as there were only limited samples for models' learning in labelling high-risk posts, which are our priority with potentially being in emergent need of clinical help [37]. Models for suicide prediction can be of greater clinical use after addressing the scarcity and imbalance of existing dataset.

With these considerations in mind, this competition raised community awareness of current obstacles in the domain of social media suicide detection and call for effective solutions. We have formulated a task for research community participation. Given 2000 posts (500 labeled posts and 1500 unlabeled posts) from r/SuicideWatch subreddits, each participating team is required to develop a predictive model that can accurately detect the corresponding suicide risk labels of the posts (i.e., indicator, ideation, behavior, or attempt). The design of the provided training data posed challenges for teams in boosting model performance while managing issues of overfitting, scarce ground-truth labels, and imbalanced risk distribution in datasets. This overview paper makes the following contributions: a). we introduce a significant task of suicide risk detection on social media posts aiming at addressing the technical and clinical deficiencies in this field; b). we highlight the utility of the 4-label suicide risk annotation scheme and a fine-grained benchmark suicide dataset; c). we provide a detailed summarization of the approaches developed by participating teams, our evaluation methods and an overview of the results.

II. Task Definition and Submissions

The main task for this challenge involves detecting suicide risk level given the textual information of social media posts. In particular, participating teams are required to develop a predictive model that can accurately classify the posts into the four suicide risk levels (i.e., indicator, ideation, behavior, attempt) based on the text content of posts. Each level of suicide risk is defined following the adapted version of Columbia-Suicide Severity Rating Scale [32] (Table I).

TABLE I: Detailed definition of different suicide risk levels.

Risk level	**Definitions**
Indicator (IN)	The post content has no explicit suicidal expression or has explicit expression of resolved suicidal issues.
Ideation (ID)	The post content has explicit suicidal expression but there is no plan or tendency to commit suicide.
Behaviour (BR)	The post content has explicit suicidal expression and there is plan or tendency to act out self-harm or suicide, or mentioned historical experience of self-harm behaviour.
Attempt (AT)	The post content has explicit expression concerning recent suicide attempt, or mentioned historical experience of suicide attempt.

The text content from a dataset containing 2,000 Reddit posts (500 labeled posts with suicide risk levels and 1,500 unlabeled posts), is provided to the team as training data to build the predictive model. The risk level distributions among the 500 labeled posts are N=129 for Indicator (IN), 200 for Ideation (ID), 132 for behavior (BR), and 39 for Attempt (AT). Within the limited sample size, it can be observed the dominating proportions of ID and BR while the proportion of AT category is the smallest, which suggested an imbalanced dataset revealing the clinical prevalence of lower suicide risk level [38]. With an extra 1500 unlabeled posts, we aim to call for creative ideas among the participants in overcoming the obstacles faced by current suicide detection effort (e.g., scarce ground-truth data and uneven risk distribution) while exploring the effectiveness of various model architectures. The test datasets consist of 200 new posts with labeled suicide risk levels. Each developed model undergoes primary evaluation with 100 test posts and then final evaluation using another 100 new posts. The statistics of the dataset are shown in Table II.

TABLE II: Distribution of suicide risk levels on training and test dataset.

Labels	**Training**		**Test**	
	Labeled	Unlabeled	Primary evaluation	Final evaluation
IN	190	NA	26	26
ID	140		38	38
BR	139		28	28
AT	41		8	8
Total	500	1500	100	100

This challenge open submissions from Jun 10, 2024 to Aug 31, 2024[2]. The submitted file should report the prediction results of the 100 posts in test dataset and structured as a .xlsx file containing columns for indexing, predicted suicide risk label, and the associated probability distribution for four risk labels. Each team is permitted to have multiple submissions and each submission will be evaluated using weighted F1-score upon receiving. During the competition period, evaluation of the submission results is updated via the leaderboard on a daily basis for participating teams to keep track of their model performance. As submission period closed, each team is required to hand in the source code and a report describing their model building approach for final evaluation.

The organizing committee conducted final evaluation during Sept 1-15, 2024, based on a new set of 100 posts with labeled suicide risk level. The performance of the team's work will be scored considering the three aspects: Model performance (30%), Approach innovation (40%), and Report quality (30%). Following this evaluation, the final top 10 teams are invited to submit papers for the conference and in particular, the top 3 teams in model performance are awarded extra cash prizes.

III. Data Source

All our training (N=2000) and testing (N=200) data are posts extracted from the dataset constructed by [32]. Using Reddit application programming interface (API)[3], the original dataset consists of 139,455 posts from 76,186 users published in r/SuicideWatch subreddit section between 01/01/2020 and 31/12/2021. Among these, a random sample of 500 posts were selected for suicide risk level annotations as demonstrated in [32]. We constructed our training dataset by utilizing the 500 labeled posts and randomly sampling another 1500 posts from the original dataset without annotation. Within the remaining posts of the original dataset, we ran two separate random sampling procedures, each for 100 posts, to build our test data for primary and final evaluation. The annotation of these 200 posts as test data was conducted by experienced members in the organizing committee following the same scheme proposed by [32]. To protect the privacy and confidentiality of users, all the included posts in training and test dataset were de-identified by removing users' identifying information (e.g., name, gender, age, address, and links) from post content prior to analysis.

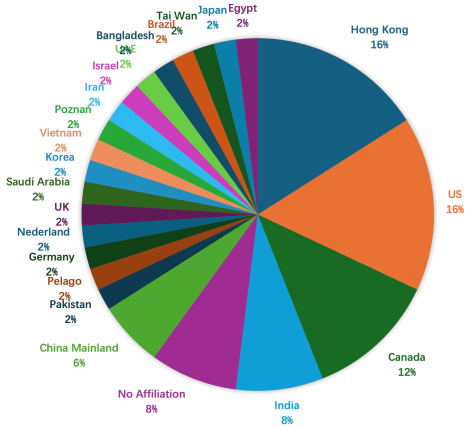

Fig. 1: Distribution of registered teams by country or area.

IV. Participating Process and Teams

A. Invitation and Registration Process

Invitation letters and posters were created to promote the competition through the following four channels:

- We displayed posters across various university campuses in Hong Kong, inviting local students to participate.
- We utilized common mailing lists such as DBWORLD, hpc-announce, and sigcomm-members@acm to share competition announcements within the research community.
- We published competition invitations through social media platforms including Twitter, WeChat, Facebook, and WhatsApp.
- We reviewed relevant publications in the field, collected contact information of the authors, and sent personalized competition invitation emails.

For the registration, a dedicated competition website were developed where users can register their teams by submitting team information and contact details through this platform after agreeing to the data usage agreement. The full text of the data usage agreement is provided in Appendix A. Upon confirming the acceptance of the data usage agreement, we send the competition data to the registered teams.

B. Participating Teams

There are 47 participating teams across 22 countries or areas shown in the Fig. 1. Among these, 21 teams submit their

[2] https://competitionpolyu.github.io/
[3] https://www.reddit.com/dev/api/

primary results on the leaderboard. 13 teams submit their final solution including one report and source code. For the final evaluation, the score of teams is assigned according to the model performance evaluated by weighted F1-score. We also invited top 10 teams to submit papers to the IEEE BigData conference according to the quality of report, innovation and model performance.

V. Results

A. *Overview of final evaluation*

The final evaluation focus on evaluate the model performance and utilize the weighted F1-score which takes the proportion of four suicide risk labels into consideration. The top 3 performing teams will be award cash prizes. Table III shows the rank of participants.

TABLE III: Final scores of teams in model performance evaluation.

Rank	Team name	Final score
1	Detection of Suicide	0.7605
2	kubapok	0.7551
3	mukumuku	0.7505
4	BioNLP@WCM	0.7463
5	Calculators	0.7341
6	The Dual	0.7312
7	BNU AI and Mental Health	0.7108
8	MindFlow	0.7072
9	EEEAT	0.6989
10	MIDAS	0.6983
11	PotatoTomato	0.6915
12	LifeWatcher	0.5528
13	Data Science and Decision Making Lab BGU	0.5496

In addition, we jointly consider the report and model performance for paper invitation. The overall score is obtained according to the following selection criteria:

- Model performance (30%)
- Approach innovation (40%)
- Report quality (30%)

Table IV display the overall score of submission. We invite the top 10 performing teams to submit papers to the conference.

TABLE IV: Overall score ranking for paper invitations.

Rank	Team name	Overall score
1	The Dual	75.436
2	BioNLP@WCM	74.389
3	Detection of Suicide	71.315
4	mukumuku	71.015
5	Calculators	70.523
6	EEEAT	67.967
7	BNU AI and Mental Health	66.824
8	LifeWatcher	65.584
9	kubapok	64.653
10	MIDAS	62.449
11	MindFlow	55.216
12	Data Science and Decision Making Lab BGU	53.488
13	PotatoTomato	48.745

TABLE V: Team strategies for addressing limited data size and suicide risk imbalance.

Team	Model	Label scarcity solutions	Label imbalance solutions
MindFlow	RoBERTa	None	Loss function
Kubapok	DeBERTa, GPT-4o	None	None
mukumuku	mental-longformer, MentaLLaMa	Pseudo-labels, Rephrasing	Loss function
D.O.S[a]	RoBERTa	Annotation, Data Generation.	Sampling
DSDM Lab BGU[b]	Phi 3.5, RoBERTa	Pseudo-labels (RoBERTa)	None
EEEAT	Qwen2-max, Claude-3.5	Pseudo-labels (bloomz-3b)	Back-translation
The Dual	Qwen2-72B-Instruct	Pseudo-labels (Llama3, DeBERTa, Qwen2)	Loss function
PotatoTomato	RoBERTa	Pseudo-labeling	None
LifeWatcher	BERT	Annotation	Oversampling
Calculators	BERT, RoBERTa, DeBERTa	None	None
BNU AI[c]	XGBoost, LSTM+Att., RoBERTa, Llama, Gemma	None	None
BioNLP@WCM	RoBERTa	Pseudo-labels (RoBERTa)	Sampling(SCS), Loss function(ICF)
MIDAS	RoBERTa	Pseudo-labels (SVM)	Data generation

[a] Detection of Suicide
[b] Data Science and Decision Making Lab BGU
[c] BNU AI and Mental Health

B. *Analysis of Solutions*

The contest of suicide ideation detection on social media attracted a diverse range of solutions from participating teams, each achieving varying levels of success. To provide a comprehensive understanding of the competition outcomes, we present an overview that summarizes key findings regarding their approaches. This overview aims to address the following questions:

- RQ 1: How to train a effective suicidal ideation model under limited data size (the 500 limited labeled posts and 1500 unlabeled posts) effectively.
- RQ 2: How to train a robust model on the dataset where the distribution of suicide risk label is imbalanced.
- RQ 3: Any conclusion on the performance between BLMs (Bert, RoBERTa, Deberta, etc.) and LLMs in this task.

Faced with challenges of limited data and suicide risk imbalance (RQ 1 and RQ 2), each participating team employed unique strategies. Table V categorizes these methods and provides an overview of the approaches utilized by each team, including their chosen models, solutions for label scarcity, and strategies to address label imbalance.

TABLE VI: Model selection and optimization.

Model type	Number of teams	Main models
BLMs	11	BERT, RoBERTa, DeBERTa, mental-longformer, SVM…
LLMs	5	Phi 3.5, Claude-3.5-Sonnet, bloomz-3b, Qwen2-72B-Instruct, Qwen2-max, Llama 3-8B, Llama 3.1-8B, Gemma 2-9B, Gemma 2-27B, GPT-4, GPT-4o, GPT-4-mini, GPT-4-turbo

TABLE VII: Prompt engineering and finetune used by competition teams.

Category	Model
Prompt engineering (in-context learning)	Qwen2-72B-Instruct, Qwen2-max, Claude-3.5-Sonnet, GPT-4, GPT-4o, GPT-4-mini, GPT-4-turbo
Fine-tuning	Llama 3-8B, Llama 3.1-8B, Gemma 2-9B, Gemma 2-27B, GPT-4o

In terms of model selection (see Table V and Table VI), solutions predominantly relied on BERT-based models among BLMs. In addition, there is a notable trend towards using LLMs as backbone models. Among these, open-source models like Llama 3 and Gemma 2 were widely adopted, while GPT-4 dominated among commercial models. Participating teams primarily employed two strategies to fine-tune LLMs: in-context learning and fine-tuning. Table VII lists the LLMs used in the competition corresponding to these two strategies. Notably, among all the LLMs, the ensemble approach incorporating fine-tuned GPT-4o achieved the best performance.

To address label scarcity, the most common approach among participants was to generate pseudo-labels for the 1,500 unlabeled posts, select high-quality pseudo-labeled samples, and then incorporate these into the training dataset. Teams proposed various techniques to enhance pseudo-label accuracy, including selecting high confidence predictions or manual verification. For label imbalance solutions, many teams employed data augmentation to supplement the minority class ('attempt') with generated posts, using approaches such as paraphrasing by LLMs and back-translation. Some teams experimented with under-sampling or over-sampling strategies to balance the distribution of risk labels before model training. To mitigate overfitting result from limited data, ensemble methods were employed by many teams to enhance robustness and maintain high performance across evaluations.

Table VIII summarizes the techniques used by teams, highlighting that data generation, ensemble learning, and loss function optimization were the most commonly employed methods. The table reveals that most teams favored a combination of approaches, reflecting both the task's complexity and the teams' innovative spirit in pursuing optimal solutions.

TABLE VIII: Overview of approaches used by teams.

Teams	Data gen.	Ensemble learn.	Class weight	Over/Under sampling	Loss func.
MindFlow					✓
Kubapok					
mukumuku	✓	✓			✓
D.O.S[a]	✓		✓	✓	✓
DSDM Lab BGU[b]					
EEEAT					
The Dual		✓			✓
PotatoTomato					
LifeWatcher		✓		✓	
Calculators					
BNU AI[c]		✓			
BioNLP@WCM		✓	✓	✓	✓
MIDAS	✓				

[a] Detection of Suicide
[b] Data Science and Decision Making Lab BGU
[c] BNU AI and Mental Health

In this competition, the final evaluation revealed some interesting insights into two type of models performance. Based on the teams' performance, we can draw a preliminary conclusion about the effectiveness of them on this task (RQ 3):

- The top ranking model is RoBERTa-based model, demonstrating its robust and effective performance in the task of suicide ideation detection. Following closely is the approach ensemble with fine-tuned GPT-4o, showcasing the potential of LLMs in this task.
- While LLMs demonstrate strong performance, their scope is relatively limited for innovation and performance improvement, primarily confined to two strategies: in-context learning and fine-tuning. The ensemble approaches combining BLMs (particularly RoBERTa) and LLMs in the contest demonstrated superior performance. For example, team BNU AI and Mental Health, which implemented aggregated models, achieved an improvement of 3.75% in weighted F1-score compared to the average performance of 3 single-model approaches.
- Regardless of the type of model used, innovative improvement designs are required to achieve excellent results. This shows that in this task, method innovation is as important as model selection.

Through comparative analysis of different teams' approaches, we identified several key technical components that contributed significantly to model effectiveness:
1) Data Enhancement Strategies: a) Teams addressing data scarcity through pseudo-labeling of the 1500 unlabeled posts is the most effective way to enhance the performance because it leverages the large amount of unlabeled data to expand the training set in a semantically meaningful way. b) Solutions employing data augmentation techniques such as back-translation demonstrated more robust performance because it introduces

linguistic variations while preserving the essential meaning of suicide-related expressions.

2) Class Imbalance Solutions: Teams implementing weighted loss functions showed better performance on minority classes, especially the "attempt" category. Specifically, weighted loss functions force the model to pay more attention to underrepresented classes such as "attempt", and this adaptive focusing mechanism helps prevent the model from being dominated by easy majority class samples during training.

The effectiveness of these components is evident in the performance patterns across teams. For instance, 2 of the top 3 teams incorporated specialized data enhancement strategies and class imbalance handling. This suggests that comprehensive technical solutions addressing multiple challenges simultaneously were more successful than those focusing on single aspects.

In general, in the competition, faced with the challenges of limited data size and label imbalance in the dataset, the teams proposed various approaches to address these issues. These results highlight the complexity of the task and reflect the innovative capability of the participating teams in tackling this challenge.

VI. Future Outlook

A. Multimodal Datasets

Even though most suicidal ideation detection tasks focus primarily on text data, the emergence of smart wearable devices such as smart bracelets, smartwatches, and AR glasses has provied new channels for suicidal ideation detection on multimodal data. These devices can monitor various physical, physiological and behavioral signals, including heart rate, blood pressure, sleep quality, physical activity, and even stress levels. In the future, the increasing availability of real-time sensor monitoring data will significantly advance our capabilities in mental health care. It will not only achieve real-time suicide risk detection but also enable the prediction of suicidal ideation.

B. Cross-linguistic Detection

The majority of current suicide detection research is predominantly focused on the English-speaking world, primarily due to the comprehensive nature of English language data, which facilitates model training. However, this focus limits the widespread application of suicide detection models across diverse regions, particularly in less developed areas where English is not the primary language. Consequently, a significant number of at-risk individuals in these regions may not receive timely detection and intervention. To address this issue, it is crucial to consider both the commonalities and differences among languages in the context of suicide detection. Two main approaches have emerged:

- Specialization in low resource languages: Some research has focused on suicide detection in low-resource languages, such as Hebrew [39] and Korean [40]. These studies aim to develop targeted solutions for specific linguistic communities, addressing the unique challenges and nuances of each language.
- Cross-lingual suicide detection: This approach seeks to create more universally applicable models, capable of detecting suicidal ideation across linguistic boundaries. While there are relatively few studies on it, the emergence of LLMs has opened new possibilities in this field due to its linguistic generalization capabilities.

In the future, the best way to detect suicidal tendency across languages may be to balance between language-specific intricacies and cross-lingual universality.

C. User Interaction and Temporal Analysis

This competition focuses on post-level suicide detection, which lacks consideration of users' background information. User-level suicide detection emphasizes historical data (including past posts and user interactions) and the temporal attributes of this data. Effectively utilizing inter-user interactions and temporal information is crucial for suicide detection. Regarding interaction information, existing works have employed social context information such as comments tree [41] and social network [42]. For temporal aspects, current research includes temporal symptom-aware attention [43], time-and phase-aware framework [44], temporal context [42]. In the future, suicide detection will continue to expand in these areas, uncovering more potential value. For instance, the impacts of user interactions are often intertwined; unraveling these complexities to identify disentangled influences will be vital for suicide ideation detection and suicide intervention. This approach holds significant value for enhancing both detection performance and the effectiveness of suicide intervention.

D. Interpretability

Automated suicide detection systems are fundamentally based on human-computer interaction, and developing interpretable model is crucial for gaining expert trust. Currently, existing models offer limited explanations for their suicide detection processes. To address this, we can approach suicide explanations from two key perspectives:

First, understanding the underlying causes of suicide. This involves models identifying suicide risk factors associated with various levels of suicidality. However, suicide often results from complex interactions between multiple risk factors, making it challenging to disentangle these relationships and extract valuable insights. The ability to clarify these connections and isolate significant suicide risk factors presents a considerable challenge in the field. Second, understanding which aspects of the input trigger the model's decision. By exploring the model's decision space to identify key information capable of altering the model's outcome, we can pinpoint the critical keywords or features influencing the decision-making process. This approach provides insight into the specific expression in the posts that drive the model's predictions, thereby enhancing the interpretability and transparency of the suicide detection system. By focusing on these aspects, we can work to develop more transparent, reliable, and effective suicide detection models.

When evaluating suicide risk detection models, we should not solely focus on the model's predictive accuracy. Instead, we must comprehensively assess the rationality of its decision-making process and the validity of its underlying reasoning. It is crucial to establish evaluation metrics for the decision rationale, ensuring that these models not only excel in experimental settings but also prove effective in the complex real-world environment.

E. Suicide Intervention

Recent advances in deep learning have significantly improved suicide risk detection models. These approaches can effectively identify individuals at risk from their posts on social media. However, effective intervention remains a critical challenge.

Globally, more and more suicide crisis intervention agencies are facing severe resource shortages. It is critical to find a way that can improve quality and efficiency of service under limited sources.

A scenario is a system can enhance human expertise by providing personality-adaptive prompts and real-time information to crisis interveners. This system considers factors like language preferences, cultural taboos, and individual's background. By combining its assistance with human judgment, interventions can become more personalized and effective, potentially improving outcomes for at-risk individuals while addressing resource limitations in suicide prevention efforts.

VII. Ethics

This challenge task involves the development of predictive models of suicide risk level among social media posts in addressing defects of available suicide dataset (e.g., limited ground-truth data, uneven risk distribution, and implicit expressions). The design and implementation of the challenge follows the ethical guidelines and suggestions proposed by [45] and [46]. The data used in this challenge was derived from an existing and publicly available dataset of Reddit posts. Considering the nature of this work, an exemption from ethical review was secured by the Institutional Review Board (IRB) of hosting organisation (The Hong Kong Polytechnic University) and approval of task content was obtained from the IEEE Bigdata committee.

We took several steps to handle the ethical considerations. First, each post in the training and test data sent to the participating teams is indexed using number with no personally identifiable information (e.g., user IDs) available. Second, to protect the privacy and confidentiality of users, the included posts were de-identified prior to analysis by removing users' identifying information (e.g., name, gender, age, address, and links) from post content. Third, the participating teams is required to sign on a Data Usage Agreement before enrolment to understand their permission, restriction of data usage and acknowledgement duty. Finally, we ensure that the competition and participating teams used the posts in a purely observational and non-intrusive manner.

Despite the aims of searching for efficient predictive models of suicide risk detection, this task cautious the direct application of these models in clinical practice. We admitted the deficiency of current machine leaning algorithms to learn perfect predictive models and errors and mistakes can occur. Therefore, any prediction results cannot be taken as formal clinical diagnosis or mean to create stigmatizing labels. One significant contribution we may provide is the algorithms for ongoing monitoring of users' suicide risk status which enabled the identification of potential high-risk users. Specifically, leveraging the ML models on social media platform is expected to promote the timely referral of risky users to professional help, facilitate the clinical diagnosis by providing extra insights of the users' profiles, and enable the early healthcare intervention [45] [47].

VIII. Conclusion

The IEEE BigData 2024 Cup Challenges@Suicide Detection on Social Media competition attracted 47 research teams over the world, showcasing the potential of natural language processing and machine learning in addressing issues in suicide detection task. Participating teams tackled challenges such as data scarcity and class imbalance through innovative methods, exploring the performance of both BLMs and LLMs in the task. Future competitions will continue to drive development in this field, focusing on innovations in multimodal, cross-lingual, and real-time intervention aspects.

Appendix
Data Usage Agreement

In consideration of the promises and mutual covenants contained in this Agreement, Recipient agree to the terms and conditions below.

Article 1. Data Set and Grant of License

1.1 The Dataset has been compiled by members of The Hong Kong Polytechnic University and comprises publicly available data from Reddit with the purpose of detecting users at suicide risk.

1.2 The Hong Kong Polytechnic University grants Recipient a non-exclusive license to use the Data Set solely for not-for-profit educational and/or research purposes. Uses of the Data Set include, but are not limited, to viewing parts or the whole of the Data Set; comparing data or content from the Data Set with data or content in other data sets; verifying research results with the Data Set; and extracting any part of the Data Set for use in Recipient publications or Recipient research in accordance with the terms of this Agreement.

Article 2. Recipient Representations

2.1 Recipient represents that it is not bound by any pre-existing legal obligations or other applicable law(s) that prevent Recipient from receiving or using the Data Set.

2.2 Recipient shall provide proper citation and acknowledgement to The Hong Kong Polytechnic University as the source of the Data Set in Recipient publications, presentations or other public dissemination of work utilizing the Data Set.

@articleli2022suicide, title=Suicide risk level prediction and suicide trigger detection: A benchmark dataset, author=Li, Jun and Chen, Xinhong and Lin, Zehang and Yang, Kaiqi and Leong, Hong Va and Yu, Nancy Xiaonan and Li, Qing, journal=HKIE Transactions Hong Kong Institution of Engineers, volume=29, number=4, pages=268–282, year=2022, publisher=Taylor & Francis 2.3 Recipient shall use Data Set for non-commercial, educational and/or research purposes only.

2.4 Recipient shall provide The Hong Kong Polytechnic University with immediate notice in writing of any breach of this Agreement, and if identification of any user in the Data Set becomes known to Recipient, Recipient shall also immediately use its reasonable best efforts to mitigate any harm or damage from such breach.

Article 3. Recipient Restrictions

3.1 Recipient shall not deduce or obtain information from the Data Set that results in Recipient or any third party(ies) directly or indirectly identifying any research subjects with or without the aid of other information acquired elsewhere.

3.2 Recipient shall not use the Data Set in any way prohibited by applicable local, state or federal laws.

3.3 Recipient shall not modify the Data Set, except as allowed hereunder.

3.4 Recipient shall not transfer any part of the Data Set to any third party without prior written consent from The Hong Kong Polytechnic University.

3.5 Recipient shall not make or use the Data Set for any commercial purpose.

Acknowledgments

The authors would like to extend sincere gratitude to Chairs of IEEE BigData, all participants, and sponsors who made this competition possible. Special thanks to the members in the event cube project and those who, though not on the organizing committee, offered invaluable advice. The authors particularly grateful for the valuable advice from colleagues who previously organized competitions at Hong Kong Polytechnic University.

References

[1] W. H. Organization *et al.*, "Suicide worldwide in 2019: global health estimates," 2021.

[2] J. C. Franklin, J. D. Ribeiro, K. R. Fox, K. H. Bentley, E. M. Kleiman, X. Huang, K. M. Musacchio, A. C. Jaroszewski, B. P. Chang, and M. K. Nock, "Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research." *Psychological bulletin*, vol. 143, no. 2, p. 187, 2017.

[3] M. Corke, K. Mullin, H. Angel-Scott, S. Xia, and M. Large, "Meta-analysis of the strength of exploratory suicide prediction models; from clinicians to computers," *BJPsych open*, vol. 7, no. 1, p. e26, 2021.

[4] K. Kusuma, M. Larsen, J. C. Quiroz, M. Gillies, A. Burnett, J. Qian, and M. Torok, "The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: A meta-analysis and systematic review," *Journal of psychiatric research*, vol. 155, pp. 579–588, 2022.

[5] K. M. Schafer, G. Kennedy, A. Gallyer, and P. Resnik, "A direct comparison of theory-driven and machine learning prediction of suicide: A meta-analysis," *PloS one*, vol. 16, no. 4, p. e0249833, 2021.

[6] C. R. Cox, E. H. Moscardini, A. S. Cohen, and R. P. Tucker, "Machine learning for suicidology: A practical review of exploratory and hypothesis-driven approaches," *Clinical Psychology Review*, vol. 82, p. 101940, 2020.

[7] R. Jacobucci, A. K. Littlefield, A. J. Millner, E. M. Kleiman, and D. Steinley, "Evidence of inflated prediction performance: A commentary on machine learning and suicide research," *Clinical Psychological Science*, vol. 9, no. 1, pp. 129–134, 2021.

[8] R. C. Kessler, R. M. Bossarte, A. Luedtke, A. M. Zaslavsky, and J. R. Zubizarreta, "Suicide prediction models: a critical review of recent research with recommendations for the way forward," *Molecular psychiatry*, vol. 25, no. 1, pp. 168–179, 2020.

[9] E. M. Kleiman, C. R. Glenn, and R. T. Liu, "The use of advanced technology and statistical methods to predict and prevent suicide," *Nature reviews psychology*, vol. 2, no. 6, pp. 347–359, 2023.

[10] A. Pigoni, G. Delvecchio, N. Turtulici, D. Madonna, P. Pietrini, L. Cecchetti, and P. Brambilla, "Machine learning and the prediction of suicide in psychiatric populations: a systematic review," *Translational psychiatry*, vol. 14, no. 1, p. 140, 2024.

[11] J. G. C. Ramírez, M. M. Islam, and A. I. H. Even, "Machine learning applications in healthcare: Current trends and future prospects," *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, vol. 1, no. 1, 2024.

[12] B. Desmet and V. Hoste, "Online suicide prevention through optimised text classification," *Information Sciences*, vol. 439, pp. 61–78, 2018.

[13] A. G. Hevia, R. C. Menéndez, and D. Gayo-Avello, "Analyzing the use of existing systems for the clpsych 2019 shared task," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 148–151.

[14] S. Ji, C. P. Yu, S.-f. Fung, S. Pan, and G. Long, "Supervised learning for suicidal ideation detection in online user content," *Complexity*, vol. 2018, no. 1, p. 6157249, 2018.

[15] J. Du, Y. Zhang, J. Luo, Y. Jia, Q. Wei, C. Tao, and H. Xu, "Extracting psychiatric stressors for suicide from social media using deep learning," *BMC medical informatics and decision making*, vol. 18, pp. 77–87, 2018.

[16] M. Morales, P. Dey, T. Theisen, D. Belitz, and N. Chernova, "An investigation of deep learning systems for suicide risk assessment," in *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, 2019, pp. 177–181.

[17] R. Sawhney, P. Manchanda, P. Mathur, R. Shah, and R. Singh, "Exploring and learning suicidal ideation connotations on social media with deep learning," in *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2018, pp. 167–175.

[18] X. Zhao, S. Lin, and Z. Huang, "Text classification of micro-blog's" tree hole" based on convolutional neural network," in *Proceedings of the 2018 international conference on algorithms, computing and artificial intelligence*, 2018, pp. 1–5.

[19] T. H. Aldhyani, S. N. Alsubari, A. S. Alshebami, H. Alkahtani, and Z. A. Ahmed, "Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models," *International journal of environmental research and public health*, vol. 19, no. 19, p. 12635, 2022.

[20] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Computing and Applications*, vol. 34, no. 13, pp. 10 309–10 319, 2022.

[21] J. Li, S. Zhang, Y. Zhang, H. Lin, and J. Wang, "Multifeature fusion attention network for suicide risk assessment based on social media: algorithm development and validation," *JMIR medical informatics*, vol. 9, no. 7, p. e28227, 2021.

[22] A. Roy, K. Nikolitch, R. McGinn, S. Jinah, W. Klement, and Z. A. Kaminsky, "A machine learning approach predicts future risk to suicidal ideation from social media data," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–12, 2020.

[23] A. Abdulsalam and A. Alhothali, "Suicidal ideation detection on social media: A review of machine learning methods," *Social Network Analysis and Mining*, vol. 14, no. 1, pp. 1–16, 2024.

[24] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, "Mentallama: interpretable mental health analysis on social media with large language models," in *Proceedings of the ACM on Web Conference 2024*, 2024, pp. 4489–4500.

[25] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general artificial intelligence? a first

evaluation of chatgpt," *IEEE Intelligent Systems*, vol. 38, no. 2, pp. 15–23, 2023.

[26] B. Lamichhane, "Evaluation of chatgpt for nlp-based mental health applications," *arXiv preprint arXiv:2303.15727*, 2023.

[27] T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, and K. K. Bressem, "Medalpaca–an open-source collection of medical conversational ai models and training data," *arXiv preprint arXiv:2304.08247*, 2023.

[28] X. Xu, B. Yao, Y. Dong, H. Yu, J. Hendler, A. K. Dey, and D. Wang, "Leveraging large language models for mental health prediction via online text data," *arXiv preprint arXiv:2307.14385*, 2023.

[29] J. Chim, A. Tsakalidis, D. Gkoumas, D. Atzil-Slonim, Y. Ophir, A. Zirikly, P. Resnik, and M. Liakata, "Overview of the clpsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts," in *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, 2024, pp. 177–190.

[30] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, 2018, pp. 25–36.

[31] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. Sheth, R. Welton, and J. Pathak, "Knowledge-aware assessment of severity of suicide risk for early intervention," in *The world wide web conference*, 2019, pp. 514–525.

[32] J. Li, X. Chen, Z. Lin, K. Yang, H. V. Leong, N. X. Yu, and Q. Li, "Suicide risk level prediction and suicide trigger detection: A benchmark dataset," *HKIE Transactions Hong Kong Institution of Engineers*, vol. 29, no. 4, pp. 268–282, 2022.

[33] L. Cao, H. Zhang, L. Feng, Z. Wei, X. Wang, N. Li, and X. He, "Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention," *arXiv preprint arXiv:1910.12038*, 2019.

[34] P. P. Sinha, R. Mishra, R. Sawhney, D. Mahata, R. R. Shah, and H. Liu, "# suicidal-a multipronged approach to identify and explore suicidal ideation in twitter," in *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, pp. 941–950.

[35] E. D. Klonsky and A. M. May, "The three-step theory (3st): A new theory of suicide rooted in the "ideation-to-action" framework," *International Journal of Cognitive Therapy*, vol. 8, no. 2, pp. 114–129, 2015.

[36] Y.-K. Kim, "Frontiers in psychiatry artificial intelligence, precision medicine, and other paradigm shifts preface," *FRONTIERS IN PSYCHIATRY: ARTIFCIAL INTELLIGENCE, PRECISION MEDICINE, AND OTHER PARADIGM SHIFTS*, vol. 1192, pp. V–VII, 2019.

[37] B. E. Belsher, D. J. Smolenski, L. D. Pruitt, N. E. Bush, E. H. Beech, D. E. Workman, R. L. Morgan, D. P. Evatt, J. Tucker, and N. A. Skopp, "Prediction models for suicide attempts and deaths: a systematic review and simulation," *JAMA psychiatry*, vol. 76, no. 6, pp. 642–651, 2019.

[38] Y. Yan, J. Hou, Q. Li, and N. X. Yu, "Suicide before and during the covid-19 pandemic: a systematic review with meta-analysis," *International journal of environmental research and public health*, vol. 20, no. 4, p. 3346, 2023.

[39] A. Bialer, D. Izmaylov, A. Segal, O. Tsur, Y. Levi-Belz, and K. Gal, "Detecting suicide risk in online counseling services: A study in a low-resource language," in *COLING*. International Committee on Computational Linguistics, 2022, pp. 4241–4250.

[40] S. Park, K. Park, J. Ahn, and A. Oh, "Suicidal risk detection for military personnel," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 2523–2531. [Online]. Available: https://aclanthology.org/2020.emnlp-main.198

[41] R. Sawhney, S. Agarwal, A. T. Neerkaje, N. Aletras, P. Nakov, and L. Flek, "Towards suicide ideation detection through online conversational context," in *SIGIR*. ACM, 2022, pp. 1716–1727.

[42] R. Sawhney, H. Joshi, R. R. Shah, and L. Flek, "Suicide ideation detection via social and temporal user representations using hyperbolic learning," in *NAACL-HLT*. Association for Computational Linguistics, 2021, pp. 2176–2190.

[43] D. Lee, S. Son, H. Jeon, S. Kim, and J. Han, "Towards suicide prevention from bipolar disorder with temporal symptom-aware multitask learning," in *KDD*. ACM, 2023, pp. 4357–4369.

[44] R. Sawhney, H. Joshi, L. Flek, and R. R. Shah, "PHASE: learning emotional phase-aware representations for suicide ideation detection on social media," in *EACL*. Association for Computational Linguistics, 2021, pp. 2415–2428.

[45] A. Benton, G. Coppersmith, and M. Dredze, "Ethical research protocols for social media health research," in *EthNLP@EACL*. Association for Computational Linguistics, 2017, pp. 94–102.

[46] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. B. Silenzio, and M. D. Choudhury, "A taxonomy of ethical tensions in inferring mental health states from social media," in *FAT*. ACM, 2019, pp. 79–88.

[47] A. Zirikly, P. Resnik, Ö. Uzuner, and K. Hollingshead, "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*, K. Niederhoffer, K. Hollingshead, P. Resnik, R. Resnik, and K. Loveys, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 24–33. [Online]. Available: https://aclanthology.org/W19-3003